Prediction of Medical Charge to Help the Insurance Company Suggesting the Insurance Plan of Their Future Customer Based on The Historical Charges

Team 4

Mohammad Ridwan

Tsai Chin Yu

Ting Yeh Yang

Abstract

This paper discusses how to use a medical cost personal dataset to predict the charge so that the insurance company can assist the insured to choose the proper insurance plan based on the predicted charge. According to the whole give attributes dataset, the best regression model which works best for the charge’s prediction is the nonlinear model, which is Multivariate adaptive regression splines (MARS). From that model, we found that there are three variables, which are smoker, age, and children, will be the focus of the insurance company attention as those three attributes serves as the variable importance of the MARS model. Another finding based on the regression model result, the insurance company may help the younger future customer, which the ages are below 40 years old, to choose High-Deductible Health Plan with or Without a Health Savings Account as the premium in general is lower compared to the other plans. However, it does not rule out for suggesting the younger age to choose the insurance plan that has higher deductible price as the high bmi also found among the beneficiary below 30 years old. Other than that, in

southeast region, the smoker tends to be the younger beneficiary who might improve the probability of getting the higher charge.

To improve the model performance, adding more attribute such as the beneficiary income, hospital, and doctor specialty to the dataset, should be done in the future. Also, by providing such information, the insurance company may be able to create new type of insurance plan to adjust the needs of the beneficiary based on the behavior or region residency.

Other than that, the result of this regression model ca be used for the insurance company as well to get the profitable customer. In detail, if the customers charge amount is small, insurance company can target them sell another insurance plan. It is because the customers who have a lower claim amount tend to be healthier or safer than other customers.

Insurance

Insurance is a policy that help an individual get financial protection against losses (Kagan, Insurance, 2021). The most common type of insurance policies is auto, health, homeowners, and life (Kagan, Insurance, 2021).

Other than that, there are components in the insurance policy such as premium, policy limit, and deductible (Kagan, Insurance, 2021). In our project we will focus on the premium component as such component is the component that will be predicted to help the insurance company target or focus on the customer or insurers.

Insurance premium

Insurance premium is amount of money that insurers pay for an insurance policy (Kagan, Insurance Premium, 2020). In other words, it is paid for policies to cover healthcare, auto, home, and life insurance (Kagan, Insurance Premium, 2020).

The price of the premium depends on some factors such as age, type of coverage, area where you live in, as well as the historical claim (Kagan, Insurance Premium, 2020). For an instance, as our project focus on the health insurance charge prediction, the insurance premium can be defined as amount that should be paid every month to the health insurance company if you are covered by the health insurance (USA.gov, n.d.).

Insurance Plan

Choosing the insurance healthcare plan is not that easy as the plan with the lowest monthly premium does not guarantee to be the best fit for the customers (USA.gov, n.d.). Other than that, if the customers need much health care, a plan with a bit higher premium but a less deductible may save the customers money (USA.gov, n.d.). Therefore, in our project, we aim to help the insurance companies to help their future customers in choosing the plan based on the historical claim dataset. In addition, the prediction may also help the customers to wisely spend the amount of money in paying the insurance premium.

The insurance plan in each health insurance may be different. For an instance, there are four insurance plans in Independence Blue Cross (Independence Blue Cross, 2021). The insurance plans are such the follows (Independence Blue Cross, 2021):

- Employer-based health insurance

This type of plan is where the employer purchases insurance on behalf of the employees and may cover all or some of the cost of the plan premium. (Independence Blue Cross, 2021)

- Medicare

  This insurance plan does not see the customer income (Independence Blue Cross, 2021). However, for people 65 years or older or for those who are younger with disabilities (Independence Blue Cross, 2021).

- Medicaid

  This plan is intended for most vulnerable individuals from all backgrounds, and geographical regions who meet certain income and other eligibility requirements (Independence Blue Cross, 2021)

- Individual health insurance

  This plan can be purchased by the customer itself as the opposed of the Employer-based health insurance plan (Independence Blue Cross, 2021).

Dataset

The dataset that we used in our project was obtained through Kaggle website (Choi, 2018). Specifically, the dataset is a medical cost personal dataset which contains the following variables (Choi, 2018):

- **age**

  This variable contains nominal data. It explains the age of the primary beneficiary (Choi, 2018)

- **sex**

This variable contains categorical data. It shows the insurance contractor gender that comprises of male and female (Choi, 2018).

- **bmi**

  This variable contains nominal data. In detail, it contains the body mass index which provides the calculation of weights that are relatively high or low relative to height (Choi, 2018).

- **children**

  This variable contains nominal data. It explains the number of children which are covered by health insurance (Choi, 2018).

- **smoker**

  This variable contains categorical data that indicates whether the primary beneficiary is a smoker or not (Choi, 2018). There are two levels of data in this attribute, which are yes and no.

- **region**

  This variable has categorical data that explains the residential area of the beneficiary in the USA (Choi, 2018). It has four level of categorical such as northeast, southeast, southwest, and northwest (Choi, 2018).

- **charges**

  This variable is the nominal data which shows the amount of medical costs billed by the health insurance (Choi, 2018). Also, it serves as the Y variable that will be predicted in this project.

In addition, the charges of the medical costs will be predicted based on the other 6 variables above such as age, sex, bmi, children, smoker, region, and charges.

Other than that, we provide the sample of 10 rows data of each variable such as the follows:

```
     age sex      bmi children smoker region    charges
   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
 1    19 female  27.9        0 yes    southwest  16885.
 2    18 male    33.8        1 no     southeast   1726.
 3    28 male    33          3 no     southeast   4449.
 4    33 male    22.7        0 no     northwest  21984.
 5    32 male    28.9        0 no     northwest   3867.
 6    31 female  25.7        0 no     southeast   3757.
 7    46 female  33.4        1 no     southeast   8241.
 8    37 female  27.7        3 no     northwest   7282.
 9    37 male    29.8        2 no     northeast   6406.
10    60 female  25.8        0 no     northwest  28923.
# ... with 1,328 more rows
```

Prior to the prediction process, there are several steps of transformation that will be done to the dataset so that the performance or the model accuracy can be improved.

Data Cleansing

In this step we will do the fixing or removing incorrect, duplicate, or incomplete data within a dataset as the definition of data cleansing (Tableau Software, LLC, 2003-2001). The process in this project consists of checking the missing values, and removing duplicate values, and removing the extreme outliers.

- **Missing Values**

  As the dataset is pretty good, it does not contain any missing value in it.

```
> any_na(insurance_data)
[1] FALSE
```

- **Duplicate Values**

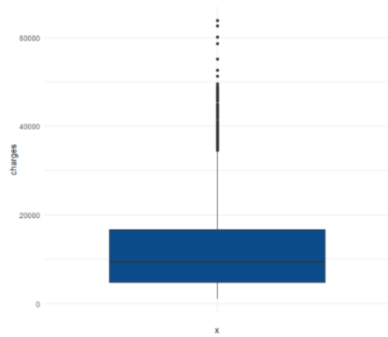
  There is one duplicate row found in the dataset. The duplicate data can be seen in the following picture:

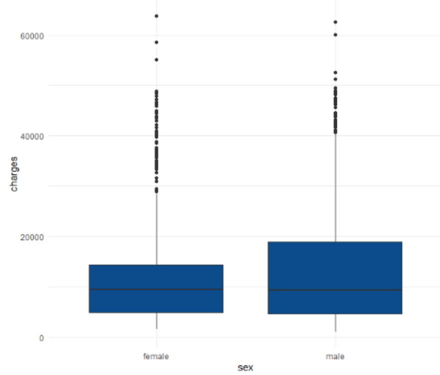
```
> #check duplicate row
> insurance_data[duplicated(insurance_data), ]
# A tibble: 1 x 7
    age sex      bmi children smoker region    charges
  <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
1    19 male    30.6        0 no     northwest   1640.
```

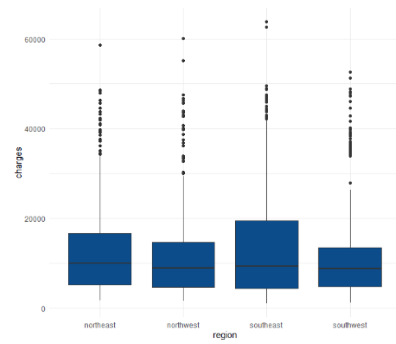
To improve the model prediction, we decided to remove the duplicate value.

- **Outliers**

The dataset contains many outliers as we can see following picture. However, removing the whole outliers might defect the model prediction. Therefore, we decided to do the further analysis so that only the extreme outliers will be removed from the dataset.







As we can see in the picture above, specifically the left side picture, there 7 extreme outliers. The boxplot on the left side above represents the charges based on the whole variables. The whole 7 extreme variables are the charges that has the amount above 50000.

To do the further analysis prior to deciding whether to omit the outliers, we look further for the charges based on the gender variables as shown in the middle side picture above. From there, we can see that the 7 outliers which were found in the previous visualization becomes more obvious to be omitted as the charges between female and male tend to be below 50000.

Other than that, we try to pay more attention to see how the outliers appears in each region as shown in the left side picture above. Further, we decided to delete the 7 outliers since such outliers look even more extreme.

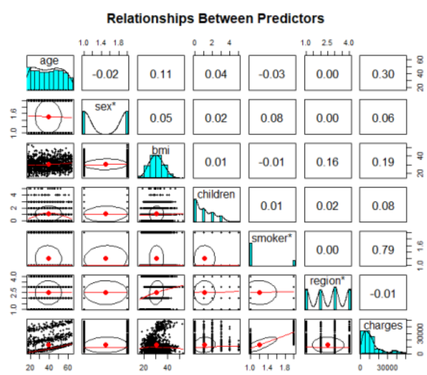
After the cleansing process, the dimension of the dataset becomes such as the follows:

```
# A tibble: 1,330 x 7
     age sex      bmi children smoker region    charges
   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
 1    19 female  27.9        0 yes    southwest  16885.
 2    18 male    33.8        1 no     southeast   1726.
 3    28 male    33          3 no     southeast   4449.
 4    33 male    22.7        0 no     northwest  21984.
 5    32 male    28.9        0 no     northwest   3867.
 6    31 female  25.7        0 no     southeast   3757.
 7    46 female  33.4        1 no     southeast   8241.
 8    37 female  27.7        3 no     northwest   7282.
 9    37 male    29.8        2 no     northeast   6406.
10    60 female  25.8        0 no     northwest  28923.
# ... with 1,320 more rows
```

In detail, the total number of rows has decreased from 1328 to 1320.

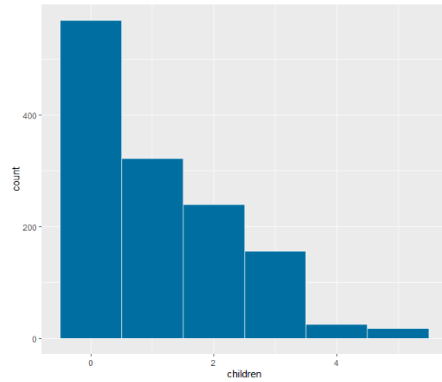
Data Exploratory

- **The matrix of scatterplots**

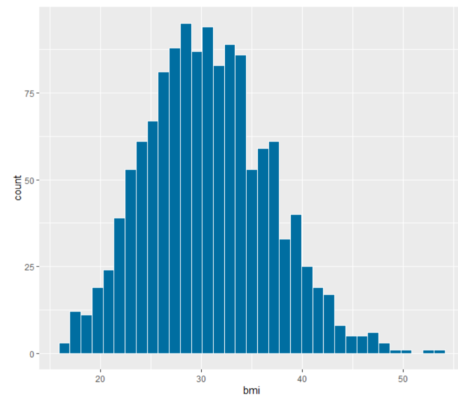


We created the matrix of scatterplots to visualize bivariate relationships between combinations of variables. Then, we find the skewness of several variables in the dataset.

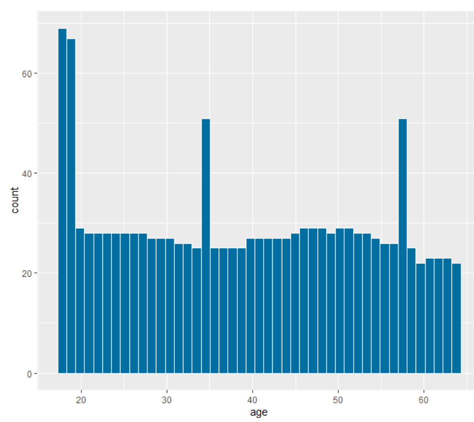
- **The skewness for the Children's data**



- **The skewness for the Bmi data**

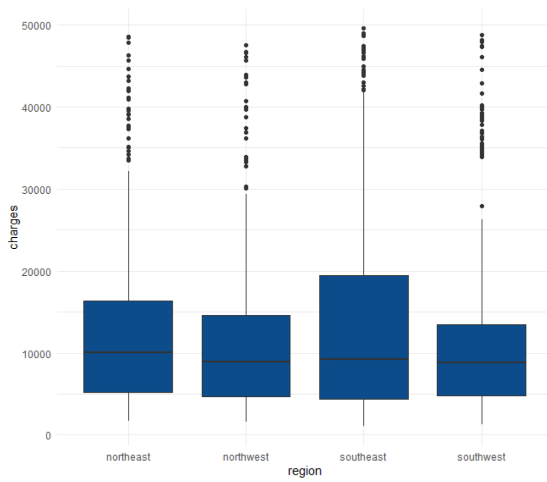


- **The skewness for the Age data**



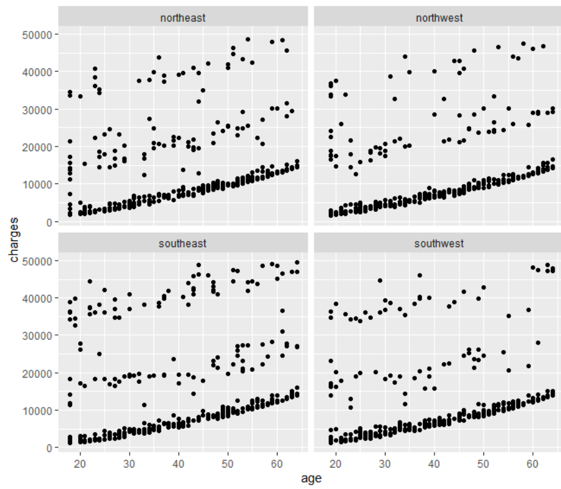
Skewness is usually described as a measure of a dataset's symmetry. The above figures show that the skewness of Children is positive skewness, the skewness of bmi close normal distribution and the skewness of age isn't obvious. We use Yeo-Johnson transformation to make our skewed data columns less skewed and more normal such that we can remove outliers. To be specific, we hope that we can make the data more normal distribution-like and improve the validity of measures of association by Yeo-Johnson transformation.

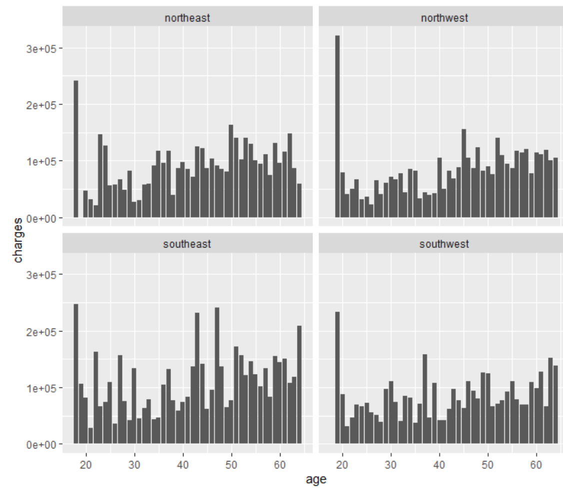
- **The Charge in each Region**



We created the relationship between region and charge. According to the above figure, the most charge range from 5000 to 20000. In our opinion, the difference charge is because the general price level is different from each region. For insurance companies, they can consider the relationship between charge and region when they make targeted charge.

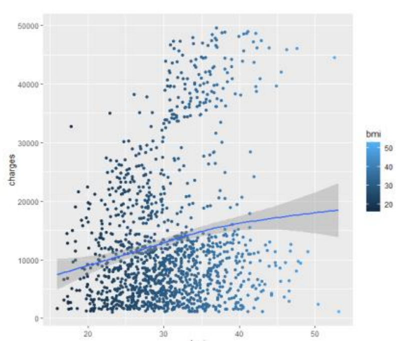
- **The Age and Charge in every Region**





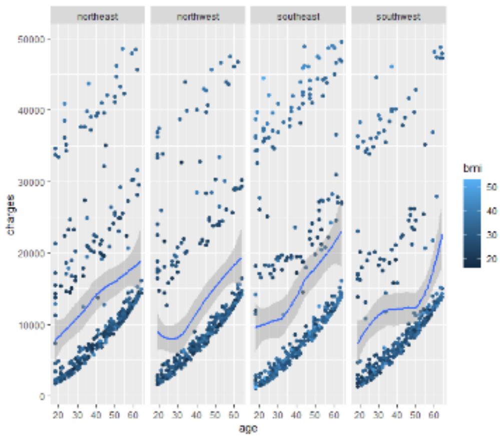
We also created the relationship between age and charge in four regions. According to the above figure, there are positive linear correlation between age and charge in every region. It shows that age can affect charge of customer obviously. For insurance companies, they can consider the relationship between charge and age when they make targeted charge.

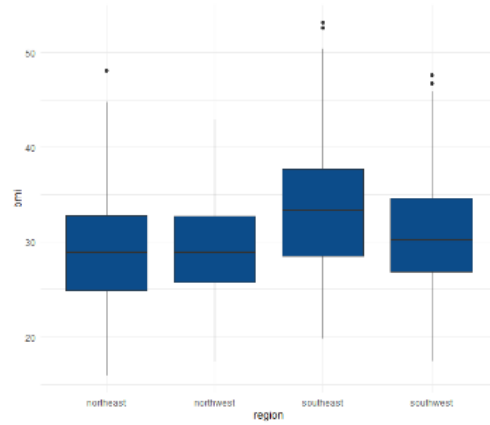
- **The relationship between Bmi and Charge**



The above figure shows the relationship between bmi and charge. There is positive linear correlation between bmi and charge. It shows that bmi can affect charge of customer obviously. For insurance companies, they can consider the relationship between charge and bmi when they make targeted charge.

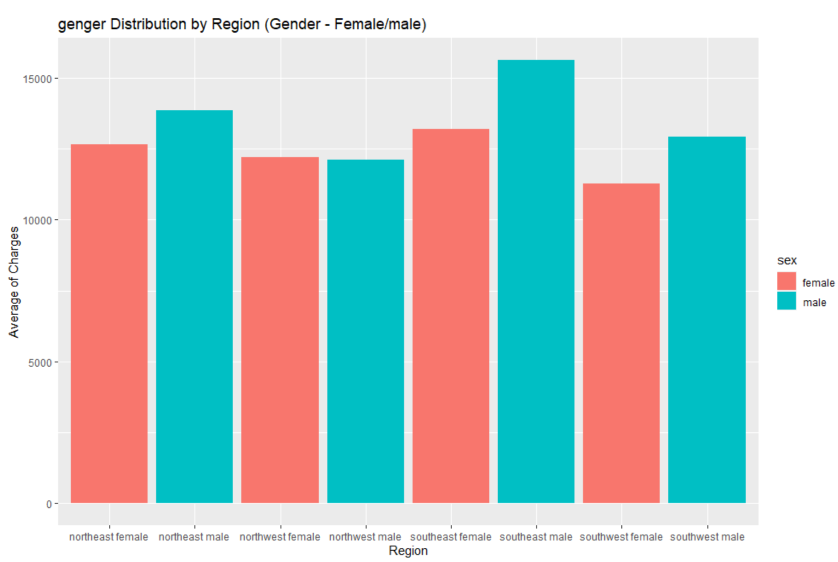
- **The compound relationship between Bmi&Age and Charge in every Region.**





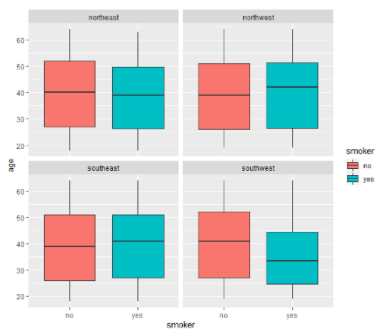
Based on the above two figures, we gained the results which age and bmi can affect the charge of customer. Therefore, we combined these two variables and region to visualize the relationship among them. In this figure, in general, the relationship between charge of customer and age& bmi are positive linear correlations. For insurance companies, this figure can help they make targeted charge by this composite result.

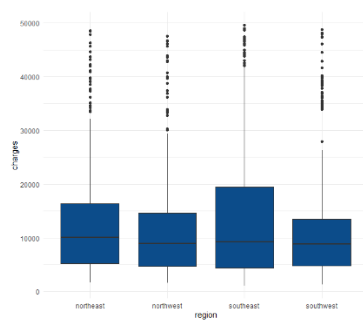
- **The comparison between Gender and Charge in every Region.**



This figure is created by gender and region. We used it to compare the charge of customer who live in different region and different gender. It presents that the charge of male is more than female in most instances. Hence, in our opinion, gender is also an important variable which can affect charge of customer. For insurance companies, they can consider the relationship between charge and gender when they make targeted charge.

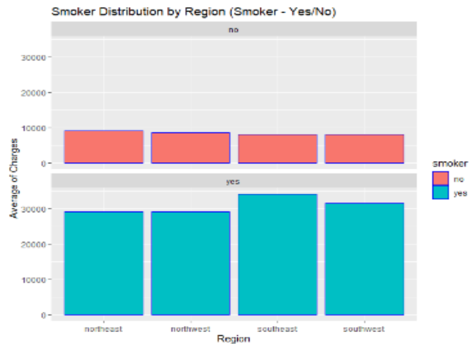
- **The comparison of Age between Smoker and Nonsmoker in every Region.**





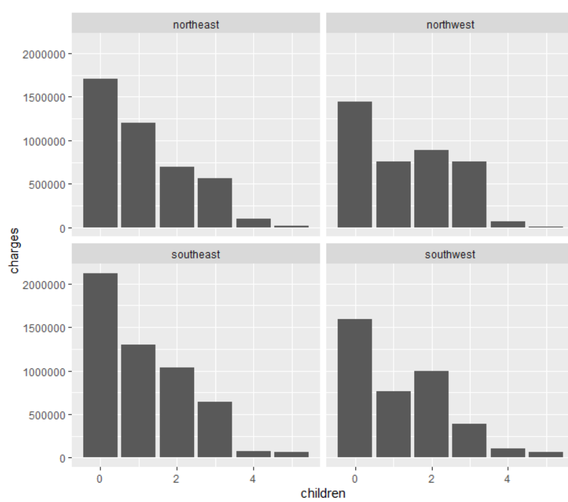
According to the first figure, we can find that the average age of smoker people who live in southeast is lower than nonsmoker people. In the second figure, the charge in the same region is higher than other regions. We guess that is related to the people who smoke tend to young in the southeast. For insurance companies, they can consider the relationship between the average age of smoker and region when they make targeted charge.

- **The comparison of Charge between Smoker and Nonsmoker in every Region.**



We created this figure to show the different charge between smoker and nonsmoker. Even though the charge of smoker is higher than nonsmoker, which is natural and right, we can find the two kinds of charge have great gaps. by visualizing. For insurance companies, they can consider the relationship between charge and gender when they make targeted charge.

- **The comparison between Charge and the number of Children in every Region**



We created this figure to show the different charge depend on the number of children in every region. Based on the results of the above figure, the charge of the family without child is higher than other families. For insurance companies, they can consider the relationship between charge and the number of children when they make targeted charge.

Data Preprocessing

- **Data Splitting**

The proportion of the splitting are 80% for train set and 20% for the test.

```
> insurance_split
<Analysis/Assess/Total>
<1066/264/1330>
```

- **Normalize**

The dataset is centered and scaled to improve the performance of the model. Afterwards, the YeoJohnson is applied to reduce the skewness of the data. However, prior to such process, we apply the recipe package so that the data transformation can be done in the train set and applied to the test set in more efficient way.

Insurance recipe

```
> insurance_recipe
Data Recipe

Inputs:

      role #variables
   outcome          1
 predictor          6
```

Apply the transformation to the Insurance recipe

```
> insurance_transformation
Data Recipe

Inputs:

      role #variables
   outcome          1
 predictor          6

Training data contained 1066 data points and no missing data.

Operations:

Centering and scaling for age, bmi, children [trained]
Yeo-Johnson transformation on age, bmi, children [trained]
```

Apply the transformation to the train set and the test set

```
> insurance_transformation_train
# A tibble: 1,066 x 7
      age sex        bmi children smoker region     charges
    <dbl> <fct>    <dbl>    <dbl> <fct>  <fct>        <dbl>
 1 -0.800 male     0.378   1.06   no     southeast    4449.
 2 -0.433 male    -1.41   -1.24   no     northwest   21984.
 3 -0.506 male    -0.299  -1.24   no     northwest    3867.
 4 -0.579 female  -0.852  -1.24   no     southeast    3757.
 5  0.488 female   0.447  -0.0725 no     southeast    8241.
 6 -0.144 female  -0.496   1.06   no     northwest    7282.
 7 -0.144 male    -0.137   0.606  no     northeast    6406.
 8  1.44  female  -0.834  -1.24   no     northwest   28923.
 9 -1.02  male    -0.766  -1.24   no     northeast    2721.
10  1.57  female  -0.753  -1.24   yes    southeast   27809.
# ... with 1,056 more rows
> insurance_transformation_test
# A tibble: 264 x 7
      age sex        bmi children smoker region     charges
    <dbl> <fct>    <dbl>    <dbl> <fct>  <fct>        <dbl>
 1 -1.47  female  -0.468   -1.24   yes    southwest   16885.
 2 -1.55  male     0.498   -0.0725 no     southeast    1726.
 3 -0.361 female   0.206   -0.0725 yes    northeast   37702.
 4  1.37  female  -0.500    1.06   no     southeast   14001.
 5 -0.579 male     0.883    0.606  yes    southwest   38711
 6 -1.25  male     0.778   -1.24   yes    southwest   35586.
 7 -1.47  female  -0.347    1.68   no     southwest    4688.
 8 -0.144 female   0.0243   0.606  no     southeast    6314.
 9  1.10  male     1.03    -1.24   no     southwest   20630.
10 -1.55  female   0.781   -1.24   no     northeast    2211.
# ... with 254 more rows
```

- **Separation between X and Y variables**

  The data separation is done since the data will be used in some models which requires the hyperparameter tuning. The purpose of this hyperparameter tuning is to improve the model performance.

X and Y data for the train set

```
> insuranceTrainX
           age    sex         bmi    children smoker    region
1  -0.800245974   male  0.37769974  1.05630804     no southeast
2  -0.433331126   male -1.41098464 -1.23807633     no northwest
3  -0.506250763   male -0.29861185 -1.23807633     no northwest
4  -0.579411833 female -0.85226151 -1.23807633     no southeast
5   0.488125333 female  0.44657787 -0.07250453     no southeast
6  -0.144308177 female -0.49633535  1.05630804     no northwest
7  -0.144308177   male -0.13711052  0.60561463     no northeast
8   1.436349712 female -0.83421527 -1.23807633     no northwest
9  -1.022932403   male -0.76586951 -1.23807633     no northeast
```

```
> insuranceTrainY
  [1]  4449.462 21984.471  3866.855  3756.622  8240.590  7281.50
 [15] 10797.336  2395.172 10602.385 36837.467 13228.847  4149.73
 [29] 15612.193  2302.300 39774.276 48173.361  3046.062  4949.75
 [43]  8059.679 47496.494 34303.167 23244.790  8606.217  4504.66
 [57] 16577.780  6799.458 11946.626 11356.661  3947.413  1532.47
 [71] 11082.577 30184.937  5729.005 47291.055  3766.884 12105.32
 [85]  3877.304  2867.120 47055.532 10825.254 11881.358  4646.75
 [99]  3385.399 17081.080  9634.538 32734.186 12815.445 13616.35
[113] 20745.989  5138.257  9877.608  1842.519  5125.216  7789.63
[127] 10450.552  5028.147  6128.797 13405.390  8116.680  1694.79
[141]  4005.423 43753.337  5325.651  4922.916 12557.605  4883.86
```

X and Y data for the test set

```
insuranceTestX
          age    sex         bmi    children smoker    region
 -1.473155466 female -0.468343581 -1.23807633    yes southwest
 -1.548760156   male  0.497896781 -0.07250453     no southeast
 -0.360664084 female  0.206245829 -0.07250453    yes northeast
  1.369648888 female -0.499839645  1.05630804     no southeast
 -0.579411833   male  0.882854910  0.60561463    yes southwest
 -1.247285752   male  0.777728206 -1.23807633    yes southwest
 -1.473155466 female -0.346796448  1.68329089     no southwest
 -0.144308177 female  0.024342723  0.60561463     no southeast
```

```
insuranceTestY
  [1] 16884.924  1725.552 37701.877 14001.134 38711.000 35585.576
 [15]  7726.854  6571.024  7935.291 37165.164  2026.974 10942.132
 [29] 27322.734 27375.905  3490.549 40720.551 10959.695 19749.383
 [43]  7731.427  3981.977  6775.961 13012.209  2483.736 25081.768
 [57] 14256.193 25992.821  1704.568 24873.385 12265.507  4349.462
 [71]  1635.734 12404.879 24603.048  1837.282 10043.249  1391.529
 [85] 46599.108 39125.332  9788.866 12638.195  5926.846  4738.268
 [99]  2201.097  1744.465 28868.664  2534.394  9880.068  1748.774
[113]  3972.925  9193.838 10923.933  6373.557 17626.240 11842.442
[127] 20149.323 13143.865  4466.621 18806.145  6435.624  5148.553
[141]  1727.540  9875.680  1263.249 16657.717 10065.413 43254.418
[155]  6933.242 27941.288 11150.780  7448.404  1917.318  2731.912
[169]  6402.291 27533.913  5458.046  8782.469  6600.361  3392.365
[183] 13462.520  6250.435 25333.333  2927.065 10096.970  9487.644
```

- **One hot Encoding (Transform to Dummy Variables)**

  This one hot encoding process transform the whole categorial variables of the X variables into dummy variables. In addition, this transformation also removes the first level of the categorial data after the transformation.

```
insuranceTrainX_dmy
          age sex.male         bmi    children smoker.yes region.northwest region.southeast region.southwest
  -0.80024597        1  0.37769974  1.05630804          0                0                1                0
  -0.43333113        1 -1.41098464 -1.23807633          0                1                0                0
  -0.50625076        1 -0.29861185 -1.23807633          0                1                0                0
  -0.57941183        0 -0.85226151 -1.23807633          0                0                1                0
   0.48812533        0  0.44657787 -0.07250453          0                0                1                0
  -0.14430818        0 -0.49633535  1.05630804          0                1                0                0
  -0.14430818        1 -0.13711052  0.60561463          0                0                0                0
   1.43634971        0 -0.83421527 -1.23807633          0                1                0                0
  -1.02293240        1 -0.76586951 -1.23807633          0                0                0                0
   1.56935761        0 -0.75331980 -1.23807633          1                0                1                0
  -1.17232598        1  0.59510722 -1.23807633          0                0                0                1
   1.16870803        0  1.39831489 -1.23807633          0                0                1                0
  -0.87427897        1  1.72661540 -1.23807633          1                0                1                0
```

```
> insuranceTestX_dmy
            age sex.male          bmi    children smoker.yes region.northwest region.southeast region.southwest
1  -1.473155466        0 -0.468343581 -1.23807633          1                0                0                1
2  -1.548760156        1  0.497896781 -0.07250453          0                0                1                0
3  -0.360664084        0  0.206245829 -0.07250453          1                0                0                0
4   1.369648888        0 -0.499839645  1.05630804          0                0                1                0
5  -0.579411833        1  0.882854910  0.60561463          1                0                0                1
6  -1.247285752        1  0.777728206 -1.23807633          1                0                0                1
7  -1.473155466        0 -0.346796448  1.68329089          0                0                0                1
8  -0.144308177        0  0.024342723  0.60561463          0                0                1                0
9   1.101432539        1  1.031412393 -1.23807633          0                0                0                1
10 -1.548760156        0  0.781499757 -1.23807633          0                0                0                0
11  0.625772639        1 -0.450887549 -0.07250453          1                0                0                1
12  1.302811335        0  0.190988617  0.60561463          0                0                0                0
13 -0.360664084        0  1.036579182  0.60561463          0                1                0                0
14  0.966406929        0 -0.433461728  1.05630804          0                0                0                1
15  0.349561969        1 -0.556066920  0.60561463          0                0                0                1
16  0.139766067        0  0.372197771 -1.23807633          0                1                0                0
17  0.418964129        0  1.176024556 -1.23807633          0                0                0                0
18 -1.247285752        1  1.078572506 -0.07250453          1                0                1                0
```

According to the transformation, we can see that there some levels that have been removed such as the non-smoker and the north east region.

Modelling

Linear Model

A linear model is a model for a continuous outcome Y of the form (ucdavis-bioinformatics-training, 2019)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

The covariate X can be continuous variables or dummy variables coding categorical covariate (ucdavis-bioinformatics-training, 2019).

The β's are unknown parameters to be estimated (ucdavis-bioinformatics-training, 2019).

The error term ϵ is assumed to be normally distributed with a variance that is constant across the range of the data (ucdavis-bioinformatics-training, 2019).

Models with all categorical covariates are referred to as ANOVA models and models with continuous covariates are referred to as linear regression models. These are all linear models, and R does not distinguish between them (ucdavis-bioinformatics-training, 2019).

In this project, there are 4 continuous variables which are age, bmi, children, and charges, and 3 categorical variables which are sex, smoker, and region.

- **Linear Regression**

  Linear regression is used to predict the value of an outcome variable *Y* based on one or more input predictor variables *X* (Prabhakaran, 2016-17). The aim is to establish a linear relationship (a mathematical formula) between the predictor variables and the response variable, so that, we can use this formula to estimate the value of the response *Y*, when only the predictors' values are known (Prabhakaran, 2016-17).

- **Robust Linear Regression**

  Robust regression is an alternative to least squares regression when data are contaminated with outliers or influential observations, and it can also be used for the purpose of detecting influential observations (Bruin, 2011).

- **Partial Least Square**

  Partial least squares regression (PLS regression) is a statistical method that bears some relation to principal components regression; instead of finding hyperplanes of minimum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space (Omics, 2016). Because both the X and Y data are projected to new spaces, the

PLS family of methods are known as bilinear factor models. Partial least squares Discriminant Analysis (PLS-DA) is a variant used when the Y is categorical (Omics, 2016).

In many data sets, the predictors we use could be correlated to the response and to each other which is not good for variability (Omics, 2016). If too many predictor variables are correlated to each other, then the variability would render the regression unstable (Omics, 2016).

PLS regression, like PCA, seeks to find components which maximize the variability of predictors but differs from PCA as PLS requires the components to have maximum correlation with the response (Omics, 2016). PLS is a supervised procedure whereas PCA is unsupervised (Omics, 2016).

- **Elastic Net**

The standard linear model (or the ordinary least squares method) performs poorly in a situation, where you have a large multivariate data set containing several variables superior to the number of samples (kassambara, 2018).

A better alternative is the penalized regression allowing to create a linear regression model that is penalized, for having too many variables in the model, by adding a constraint in the equation (James et al. 2014, P. Bruce and Bruce (2017)). This is also known as shrinkage or regularization methods (kassambara, 2018).

The consequence of imposing this penalty, is to reduce (i.e., shrink) the coefficient values towards zero (kassambara, 2018). This allows the less contributive variables to have a coefficient close to zero or equal zero (kassambara, 2018).

Elastic Net produces a regression model that is penalized with both the L1-norm and L2-norm (kassambara, 2018). The consequence of this is to effectively shrink coefficients (like in ridge regression) and to set some coefficients to zero (as in LASSO) (kassambara, 2018).

Non-linear Model

Nonlinear regression is a form of regression analysis in which data is fit to a model and then expressed as a mathematical function (Kenton, 2021). Simple linear regression relates two variables (X and Y) with a straight line (y = mx + b), while nonlinear regression relates the two variables in a nonlinear (curved) relationship (Kenton, 2021).

- **KNN**

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 neighborhood (A.Teixeira-Pinto, 2020). The size of the neighborhood needs to be set by the analyst or can be chosen using cross-validation (we will see this later) to select the size that minimizes the mean-squared error (A.Teixeira-Pinto, 2020).

While the method is quite appealing, it quickly becomes impractical when the dimension increases, i.e., when there are many independent variables (A.Teixeira-Pinto, 2020).

- **Neural Network**

Neural networks consist of simple input/output units called neurons (inspired by neurons of the human brain) (Saxena, 2020). These input/output units are interconnected, and each connection has a weight associated with it (Saxena, 2020). Neural networks are flexible and can be used for both classification and regression (Saxena, 2020).

Regression ANNs predict an output variable as a function of the inputs (Boehmke, 2018). The input features (independent variables) can be categorical or numeric types, however, for regression ANNs, we require a numeric dependent variable (Boehmke, 2018). If the output variable is a categorical variable (or binary) the ANN will function as a classifier (Boehmke, 2018).

- **Average Neural Network**

Average Neural Network is a model where the same neural network model is fit using different random number seeds (Dragićević, 2019; rdocumentation, -). All the resulting models are used for prediction (Dragićević, 2019). For regression, the output from each network is averaged (Dragićević, 2019). For classification, the model scores are first averaged, then translated to predicted classes (Dragićević, 2019).

- **Multivariate Adaptive Regression Spline**

Multivariate adaptive regression splines (MARS) provide a convenient approach to capture the nonlinearity aspect of polynomial regression by assessing cutpoints which is the same as the step functions (Boehmke, 2018). The procedure assesses each data point for each predictor as a knot and creates a linear regression model with the candidate feature(s) (Boehmke, 2018).

For evaluating the performances of all linear and non-linear regression models, we are going to measure and compare the performance by their RMSE, R-squared and MAE, which are metrics that are used for the evaluation of regression modeling.

		Train			Test		
	Model	**RMSE**	**R squared**	**MAE**	**RMSE**	**R squared**	**MAE**
Linear	lm	5822.672	0.7605086	4158.281	6094.682613	0.7088038	4253.739703
	lm log	0.437259	0.7705467	0.2802153	0.4671206	0.7363988	0.3056296
	Rlm	6651.514	0.7379445	3378.658	6909.237359	0.7004783	3627.032593
	Rlm log	**0.453711**	**0.7645428**	**0.2475506**	**0.4661475**	**0.7508068**	**0.2585461**
	pls	5822.672	0.7605086	4158.281	6094.682613	0.7088038	4253.739703
	pls log	0.437259	0.7705467	0.2802153	0.4671206	0.7363988	0.3056296
	pls tune	5817.331	0.7609104	4151.132	6083.352954	0.7098922	4246.158852
	pls tune log	0.4372475	0.7706505	0.2796326	0.4675209	0.7360181	0.3052866
	Enet	5820.738	0.7607248	4142.92	6055.299899	0.7111188	4224.020714
	Enet log	0.4372492	0.7705617	0.2802838	0.4669981	0.7365098	0.3056188
Non-Linear	KNN	6642.221	0.7115258	4016.924	5861.036012	0.7431304	3659.78815
	KNN log	0.5074329	0.7011492	0.3291472	0.4555136	0.7555373	0.2997073
	Nnet	5889.896	0.7472648	4151.438	6430.502609	0.678476	4654.393
	Nnet log	0.3758018	0.8315249	0.2077992	0.3956149	0.8100887	0.2464732
	Av Nnet	5523.432	0.7860570	3787.083	5902.077662	0.7231027	4049.111587
	Av Nnet log	0.3728731	0.8340196	0.2015077	0.3670998	0.8366974	0.2047041
	Mars	**4411.267**	**0.8594199**	**2402.391**	**4448.787071**	**0.8430872**	**2374.891851**
	Mars log	0.3949468	0.8137341	0.2151338	0.3724548	0.832324	0.214135

From the above results, with comparing the RMSE, R-squared, and MAE, we find out the best model among linear models is Robust Linear Regression, and the best model among non-linear models is Multivariate Adaptive Regression Spline. In addition, MARS is the best model among all linear and non-linear models. Thus, for predicting the insurance charges in the future, we will choose the model of Multivariate Adaptive Regression Spline.

Variable Importance

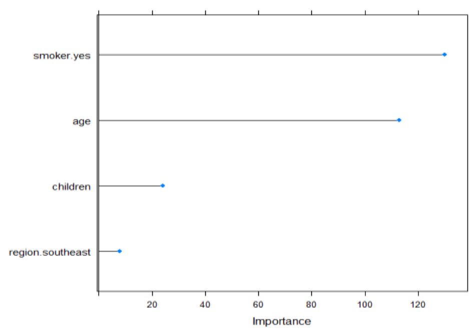
Variable importance can be referred to how much a given model used the variables to make accurate predictions (White, 2018). The more the model relies on a variable to make predictions, the more important it is for the model. It is usually used for variable selection (White, 2018).

- **Variable Importance of the Best Linear Model**

From the variable importance of the best linear model, which is Robust linear regression, we find out that people who smoke, age, number of children, and southeast region have relatively higher importance for insurance charges.

```
> rlmImp
rlm variable importance

                  Overall
smoker.yes        130.028
age               112.774
children           24.004
region.southeast    7.806
sex.male            7.721
region.southwest    7.140
bmi                 5.517
region.northwest    2.641
```

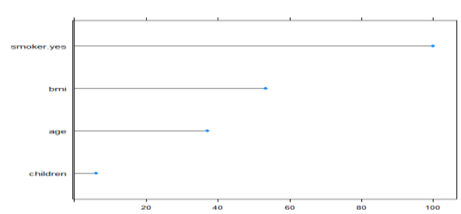


- **Variable Importance of the Best Nonlinear Model**

From the variable importance of the best non-linear model, which is Mutilvariate Adaptive Regression Spline, we find out that people who smoke, bmi, age, and number of children have relatively higher importance for insurance charges.

```
> marsImp
earth variable importance

           Overall
smoker.yes 100.000
bmi         53.294
age         37.102
children     6.025
```



Conclusion

- The best regression model which works best for the charge's prediction is the nonlinear model, which is Multivariate adaptive regression splines (MARS). It shows that the dataset that we used tends to be nonlinear as in the average performance, the nonlinear regression model performs better than the linear regression model.
- The logarithmic transformation which transforms the skewness of the charges variable only tends to improve the performance of the linear model as not the whole nonlinear model performance is improved after the transformation. In fact, the best regression model for the predicting the charges comes from the nonlinear model, which is the Multivariate adaptive regression splines (MARS).
- The variable importance between the best model of linear and nonlinear are different since the linear model tends to choose region as one of the most four important variables while the nonlinear model prefers bmi. However, the three attributes which consistently tends to affect the charges are smoker, age, and children.
- In helping the customer to choose the insurance plan or whether to improve the profit of the insurance company itself, the three variables, which are smoker, age, and children, will be the focus of the insurance company attention.
- Through this dataset, the insurance company may help the younger future customer, which the ages are below 40 years old, to choose High-Deductible Health Plan with or Without a Health Savings Account as the premium in general is lower compared to the other plans. However, it does not rule out for suggesting the younger age to choose the insurance plan that has higher deductible price as the high bmi also found among the beneficiary below

30 years old. Other than that, in southeast region, the smoker tends to be the younger beneficiary who might improve the probability of getting the higher charge.

Recommendation

- Adding more attribute such as the beneficiary income, hospital, and doctor specialty will improve the model performance as the number of row data of the dataset is quite enough.
- Some additional information such as the detail of charge such should be provided as well in the dataset as such component might help the insurance company better. Also, by providing such information, the insurance company may be able to create new type of insurance plan to adjust the needs of the beneficiary based on behavior or region residency.
- This prediction can also be used for the insurance company as well to get the profitable customer. In detail, if the customers charge amount is small, insurance company can target them sell another insurance plan. It is because the customers who have a lower claim amount tend to be healthier or safer than other customers. So, from the perspective of insurance company, if the insurance company can have healthier customers, they will not pay more as a cost for the claims. This term might indicate the loss ratio in the insurance field. Loss ratio is defined as 'claim cost' / 'premium'. If the loss ratio is higher than 100%, this means that insurance company costs more than they have received from the customers, and vice versa. So, the customers who have a lower loss ratio might will be a good customer for insurance company.

References

A.Teixeira-Pinto. (2020). *Machine Learning for Biostatistics.* Sydney, Australia: University of Sydney.

Boehmke, B. C. (2018). *UC Business Analytics R Programming Guide.* Retrieved from github: http://uc-r.github.io/ann_regression

Bruin, J. (2011, Feb). *ROBUST REGRESSION | R DATA ANALYSIS EXAMPLES*. Retrieved from newtest: https://stats.idre.ucla.edu/r/dae/robust-regression/

Choi, M. (2018, Feb 20). *Medical Cost Personal Datasets*. Retrieved from kaggle: https://www.kaggle.com/mirichoi0218/insurance

Dragićević, M. (2019, Sep 27). Is an Averaged neural network (avNNet) the average from all iterations? (Shai, Interviewer)

Independence Blue Cross. (2021). *Types of Health Insurance Plans*. Retrieved from Independence: https://www.ibx.com/htdocs/custom/getting_health_care_right/how-health-care-works/index.html#/coverage_types

Kagan, J. (2020, Aug 19). *Insurance Premium*. Retrieved from Investopedia: https://www.investopedia.com/terms/i/insurance-premium.asp

Kagan, J. (2021, March 30). *Insurance*. Retrieved from investopedia: https://www.investopedia.com/terms/i/insurance.asp

kassambara. (2018, Mar 11). *Penalized Regression Essentials: Ridge, Lasso & Elastic Net*. Retrieved from sthda: http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net/

Kenton, W. (2021, Apr 30). *Defining Nonlinear Regression*. Retrieved from investopedia: https://www.investopedia.com/terms/n/nonlinear-regression.asp

Omics, D. (2016). *Partial Least Square Regression*. Retrieved from rpubs: https://rpubs.com/omicsdata/pls

Prabhakaran, S. (2016-17). *Linear Regression*. Retrieved from r-statistics.co: http://r-statistics.co/Linear-Regression.html

rdocumentation. (-). *avNNet: Neural Networks Using Model Averaging.* Retrieved from rdocumentation: https://www.rdocumentation.org/packages/caret/versions/6.0-80/topics/avNNet

Saxena, A. (2020, Nov 10). *How Neural Networks are used for Regression in R Programming?* Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/how-neural-networks-are-used-for-regression-in-r-programming/

Tableau Software, LLC. (2003-2001). *Data cleaning: The benefits and steps to creating and using clean data*. Retrieved from tableau: https://www.tableau.com/learn/articles/what-is-data-cleaning

ucdavis-bioinformatics-training. (2019). *A Brief Introduction to Linear Models in R*. Retrieved from https://ucdavis-bioinformatics-training.github.io/2019-March-Bioinformatics-Prerequisites/thursday/linear_models.html

USA.gov. (n.d.). *Premium*. Retrieved from healthcare.gov: https://www.healthcare.gov/glossary/premium/

White, M. (2018, Mar 12). What is variable importance? (A. O, Interviewer)