6/20/2022

Machine Learning Regression Project

R and Statistical Analysis

Dr. Shadman

Mina Kanaani
Ferdowsi University Of Mashhad

Table of Contents

List of Figures

List of Tables

Machin Learning : Regression Project

Mina Kanaani

2022-06-20

Abstract

Regression Project

This is the Report on the first R and Statistical Analysis's project on regression and machine learning. In this report I aim to predict the health charges of Bangladeshi citizens based on some important features such as their age,sex, BMI, number of children they have, etc. Based on their characteristics, I built three models: Linear Regression ,KNN model and Regression Tree model. At the end, I tried out some tests to examine the prediction which showed us good results.

Why insurance cost?

Rising health care costs are a major economic and public health issue worldwide : According to the World Health Organization, health care accounted for 7.9% of Europe's gross domestic product (GDP) in 2015. In Switzerland, the health care sector contributes substantially to the national GDP, and has increased from 10.7 to 12.1% between 2010 and 2015. Moreover, because health care utilization costs may serve as a surrogate for an individual's health status, **understanding which factors contribute to increases** in health expenditures may provide insight into risk factors and potential starting points for preventive measures.

Flowchart

Using R Codes with the help of "DiagrammeR" library, I aimed to create a Flowchart to illustrate the procedure of this Project in a better way. The figure 1 below displays the project flowchart.

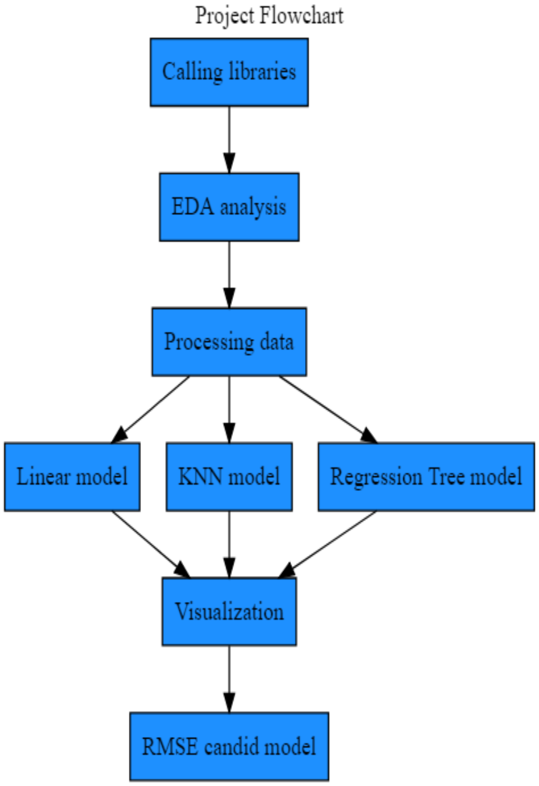

Figure 1-Project Flowchart

Recalling Libraries and Data

In this Part, we can take a short look at the data we gained from Kaggle.com, doing so, we read the excel file attached to this file to be able to access the data. Table 1 below shows the first five record of data.

age	sex	BMI	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855

Table 1- first five record of data

EDA Visualization

In this part, we use EDA analysis tools to visualize our features and their binary correlation with charges to determine which features have significant effect and which are not of much importance.

First, we check if our data has any missing values. since, there is no missing values in any features, so we can go straight to drawing plot and analysis.

In this part, First I used the simple "plot" function to draw charts of correlation between charges and age, and then charges and BMI, for both BMI and age are quantitative valuables.

First plot

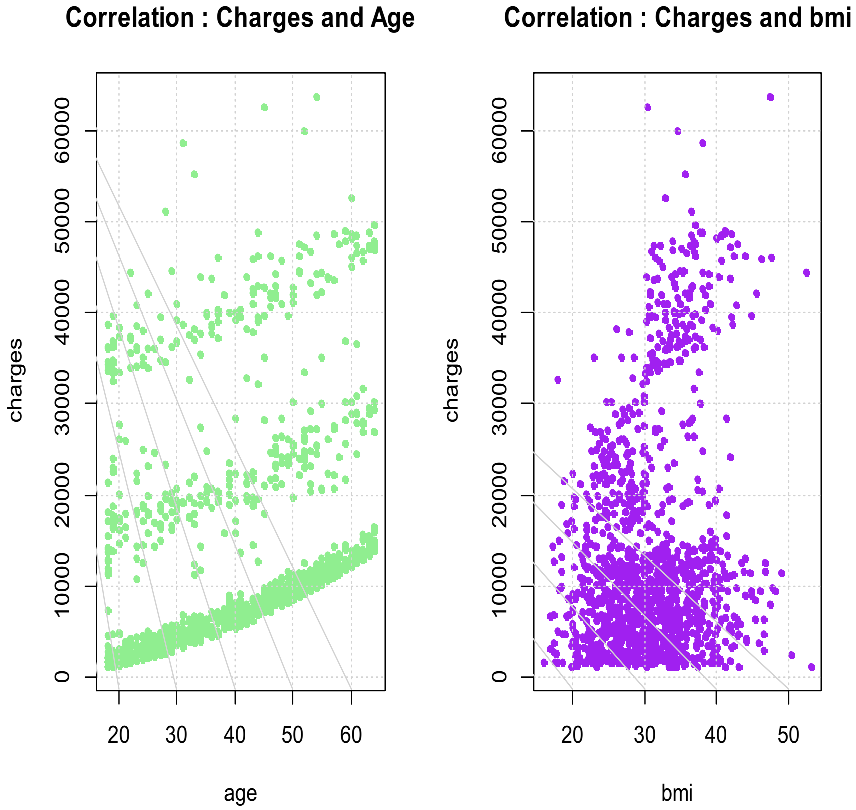

Figure 2-Correaltion between Charges-Age, and Charges-BMI

As we can see in Figure 2, as the age and BMI increase, the charges increase too. so there is an increasing trend here.

Second plot

In the next plots, I tried to plot charts for relationship between charges and sex and number of children. As it is known, the "children" and "sex" features are both qualitative variables so I cannot simply use "plot" function to draw it, for this part and next one, I use "ggplot" from package ggplot2 and with the help of jittering I reduce the amount of overlapping to make a more accurate plots.

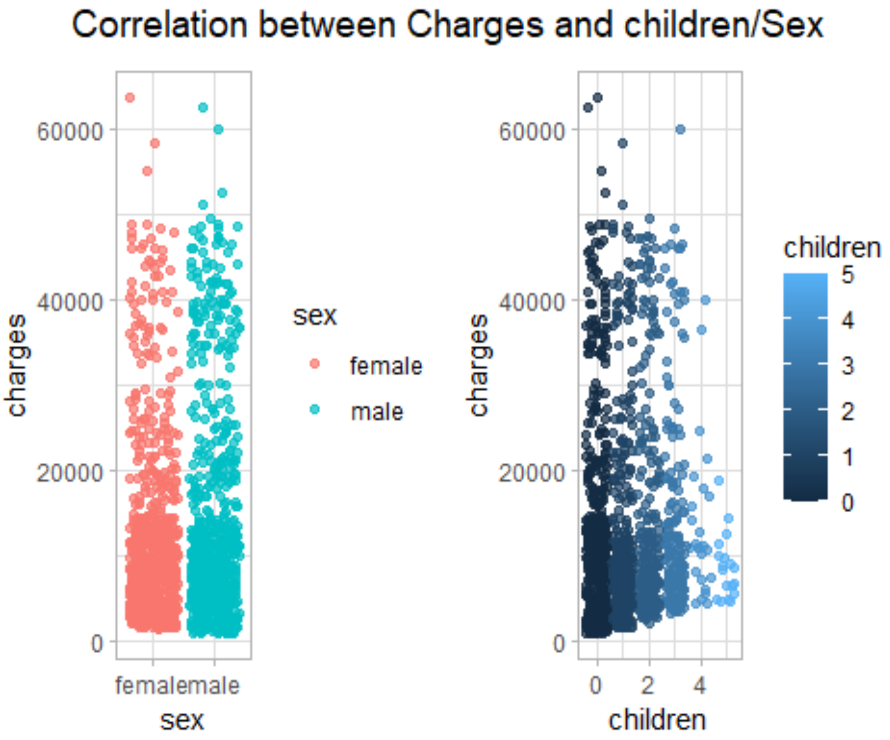

Figure 3-Correlation between Charges and Children/Sex

As illustrated in Figure 3, gender does not apparently have any effect on the charges, with number of children, we can see a decreasing trend, the more children, the less the charges which is odd and doesn't seem logical.

Third plot

In the third section, I used the same methods with ggplot and jittering to plot the charts illustrating relationship between charges with “region” and “smoker” features since they are both Qualitative.

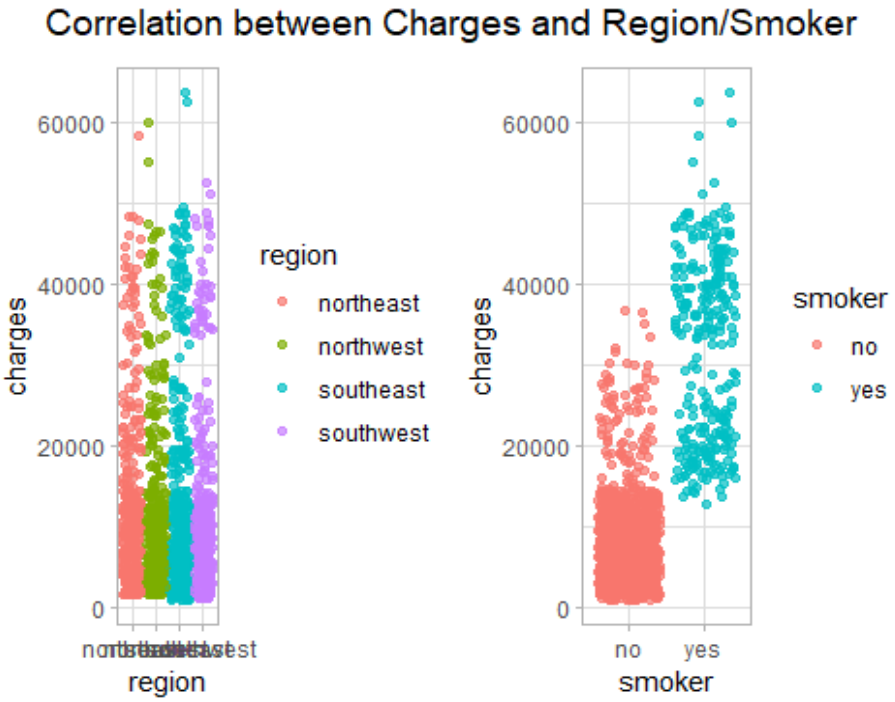

Figure 4-Correlation between Charges and Region/Smoker

As it is illustrated in Figure 4, in the result above, no obvious connection can be seen between region and charges, also as expected, charges for smokers are higher than for non_smokers.

processing data

In this part, we start of with splitting the “insu_data” into “train” and “test” . before doing so, since the analysis above illustrated , apparently “Region” and “sex” have no effect on charges so I used “subset” function to eliminated them from my main data before splitting it.

Now we start the splitting, 80% of data in train and 20% in test. I set the set.seed(42) to be able to work with constant generated numbers. After doing so, it’s time to split the train data to “estimation” and “validation”, same percentage of splitting is applicable here.

I illustrated the Figure 5 with the help of GGallay library in order to illustrate all the important features of the main Data to have a sense of their relationship.

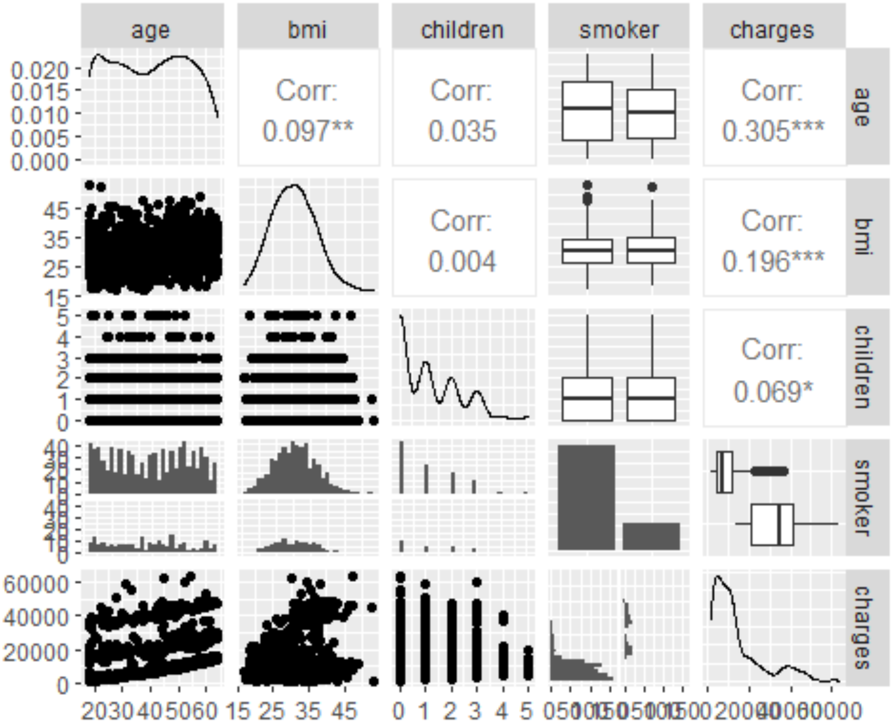

Figure 5-Important Features of the main Datae

As it is shown in the Figure 4 above, there is considerable correlation between charges with BMI and age, what we expected from the EDA visualization, we can see that the more the number of children the less is the charge which is not a logical appearance and as we expected, the smokers have more charges for insurance than non-smokers.

Linear models

In the first part of Regression modeling and fitting models to predict the charges, I used the linear modeling, using "lm" function to fit the models. The 5 models I considered for this part are consist of one that includes none of variable,one that includes 2 main quantitative variables(BMI,age), the other includes 2 main quantitative with one qualitative(smoker), the one with all of the features, and lastly, the one consists of all the variable with the binary interaction between them. Using RMSE as the main Criteria, I have compared all the models and their RMSEs for Validation data. (further Calculation and codes in R code attached to this file). Chart 2-1, illustrates the RMSE of each model.

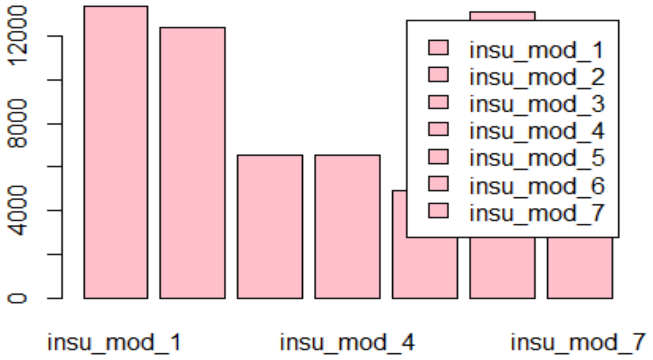

Figure 6-RMSE of each model

As we compare the validation RMSE, we can see the 5th model consist of all variables with their binary interactions with each other is the candid model with least RMSE.

So to conclude, the best model in this part is the "**insu_mod_5**" where we included all the features with their binary interaction. so we illustrate the real data with predicted using plot to see the credibility of our model. later on, we compare the other models from other methods with this one.

Visualization

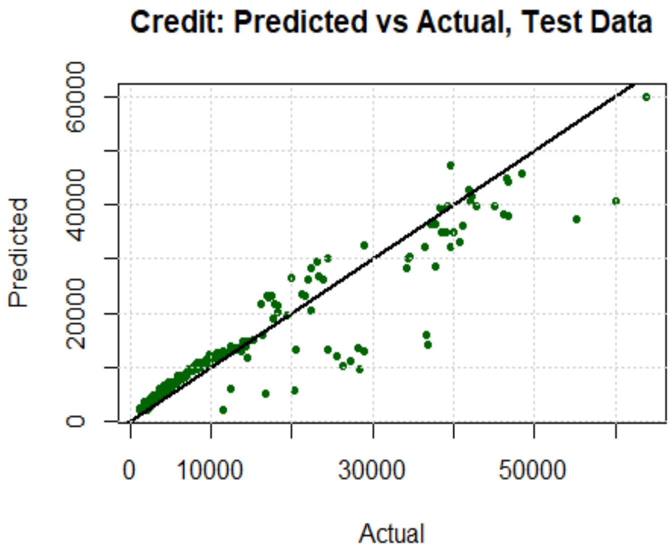

Figure 7-Credit: Predicted vs. Actual

As it is shown in Figure 7, the model mostly is moving closely to the line y=x in which x is the real "charges" and the green plot are the predicted one which at most places are near accurate.

KNN (K-Nearest Neighbor)

In the second method for creating suitable model to fit the real data and predict accordingly, I use the KNN model, in which we assume the that similar data exist in close proximity. In other words, similar data are near to each other. Before we start with fitting candidate models of KNN we have to use Scale function to normalize the quantifiable features such as age and BMI.

After we normalized our features, I created a function to generate K ranging from 1 to 100 and then, after comparing the RMSEs, we fit the final model. (Codes in the attached file)

As the Minimum RMSE shows in Table 2, The model with k=6 is the best one with lowest RMSE.

K	knn_val_rmse
1	6548.882
2	6520.816
3	6581.777
4	6499.303
5	6274.838
6	6140.336
7	6238.692

Table 2- k-values of models

Now for comparison and better analysis, we fit the models without normalization first to see the difference. We repeat all the process above with not normalized data.

K	knn_val_rmse_not
1	12673.54
2	11539.10
3	11751.59
4	11944.34
5	12029.69

Table 3- k-values of models without normalization

As it is illustrated in table 3, The model with k=2 is the best one with lowest RMSE in the not normalized features.

Now we use visualization tools to be able to compare the two normalized and not normalized models in a better way.

Visualization

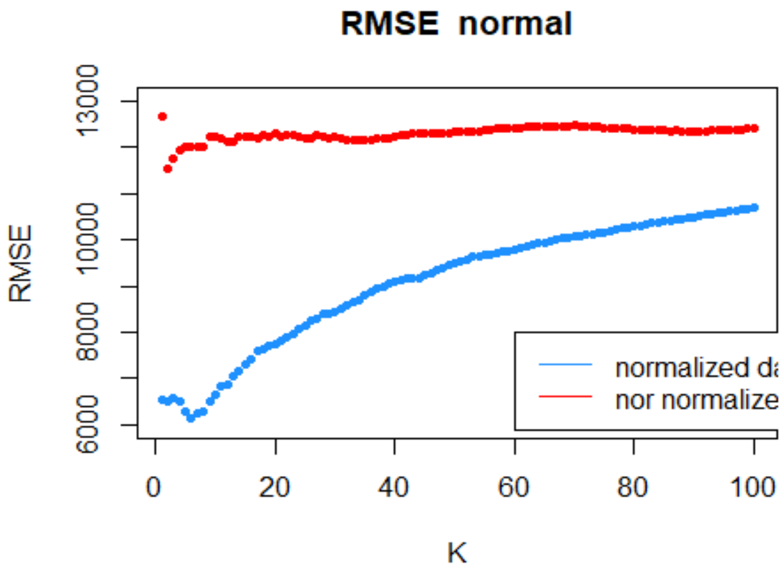

Figure 8-RMSE normal

As it is illustrated in Figure 8 above, we can see that not normalized RMSEs are higher than normalized one in any K so we can conclude that normalizing quantifiable features have significant effect on RMSE and therefore gradual candid model the best model from this method there for is 6-nearest neighbor with K=6.

So we illustrate the real data with predicted using plot to see the credibility of our model. later, we compare the other models from other methods with this one.

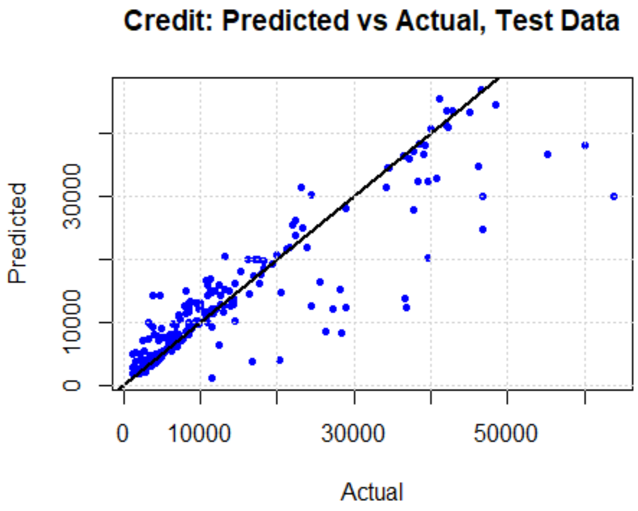

Figure 9-Predicted Vs. Actual Data

As it is shown, the model mostly is moving closely to the line y=x in which x is the real “charges” and the blue plot are the predicted one which is not as accurate as the last linear model but at most places is following the real charge.

Regression Tree

In the last method I use to fit suitable models for predicting insurance cost, I use the Regression Tree to split the data based on important features and to achieve a final node. With regression tree, I create a list with 10 models, 5 consist of minsplit=5 and cp ranging from 0 to 1. After fitting the model, we chose the one with least RMSE. According to Table 2-3, we to show each model with different attributes of cp and minsplit along with their RMSE. We can see the candid model is "insu_tree_7" with the least RMSE.

	tree_val_rmse	minsplit	cp
insu_tree_1	6079.143	20	0.000
insu_tree_2	5793.097	20	0.001
insu_tree_3	5360.543	20	0.010
insu_tree_4	6440.259	20	0.100
insu_tree_5	13374.093	20	1.000
insu_tree_6	5039.390	5	0.000
insu_tree_7	4976.345	5	0.001
insu_tree_8	5360.543	5	0.010
insu_tree_9	6440.259	5	0.100
insu_tree_10	13374.093	5	1.000

Table 4- k-values of models without normalization

Now using package “rpart”, we draw regression trees to analysis on the data given.

Visualization

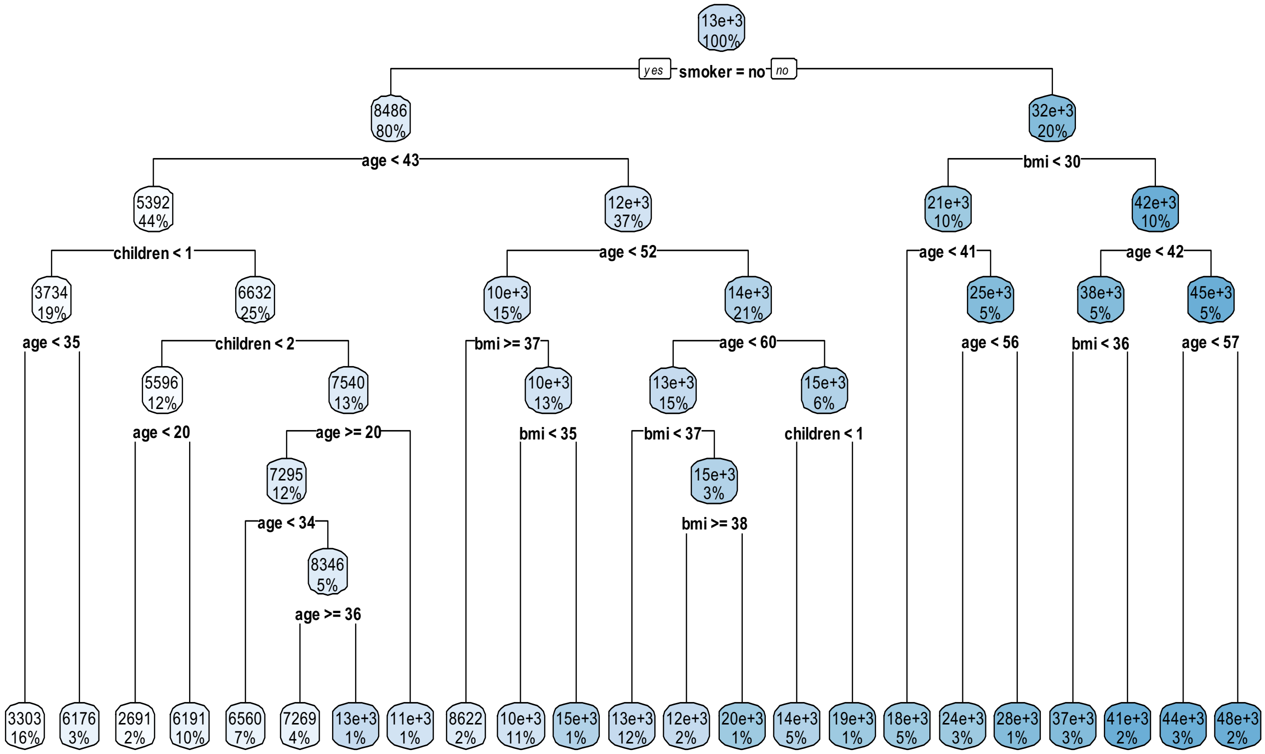

Figure 10-Visualization

As we can see in the Figure 10, most frequent cut offs are based on aged so we can say it’s our most important variables, following there are several splits with BMI variable which indicates that these two quantity variables are important. there is one first split on smoker which as we expect should have serious effect on insurance costs and splits on children which we can conclude that all 4 features are effective.

For example, from the first split, we check if the person is a smoker or not,if they are smokers, we go to the left side of the chart, now if the person age is lower that 43 ,we go to the lest part and we check the number of children they have, if it is more than 1, we go to right side and check again if the number of children is more than 2, this part we assume that the number is lower than 2, so we go to left ,there is another split on age to check whether they are under 20 or not , if they are older than 20 years old , we can predict that the insurance charge is 6191 and only 10% of all the reported data have the same conditions.

Now as the same as the previous method, we visualize our candid method to see the trend of predicted data with actual ones.

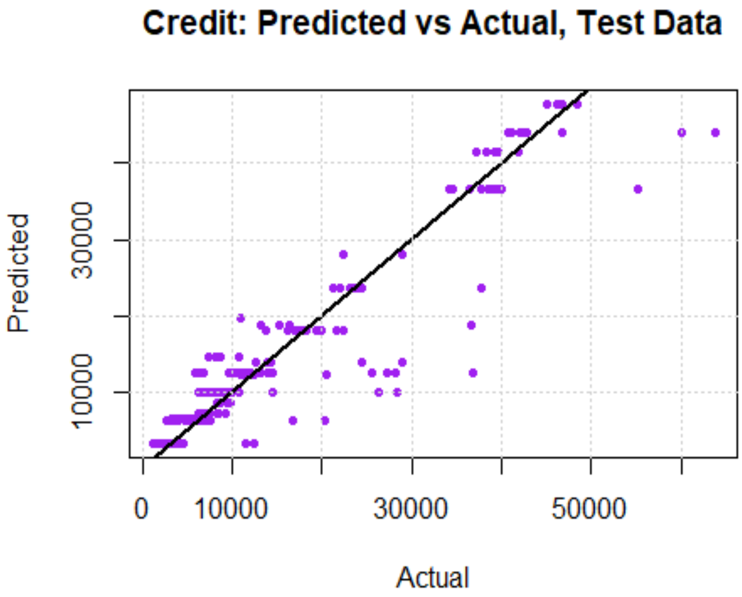

Figure 11-Predicted Vs. Actual

As it is shown in Figure 11, the model mostly is moving closely to the line y=x in which x is the real "charges" and the purple plot are the predicted one which is even more accurate than the last two models and at most places is following the real charge.

Visualization of 3 models

In the last section, we put together the tree candid models from Linear, KNN, Regression Tree method and then by calculating RMSEs we choose our final optimal model.

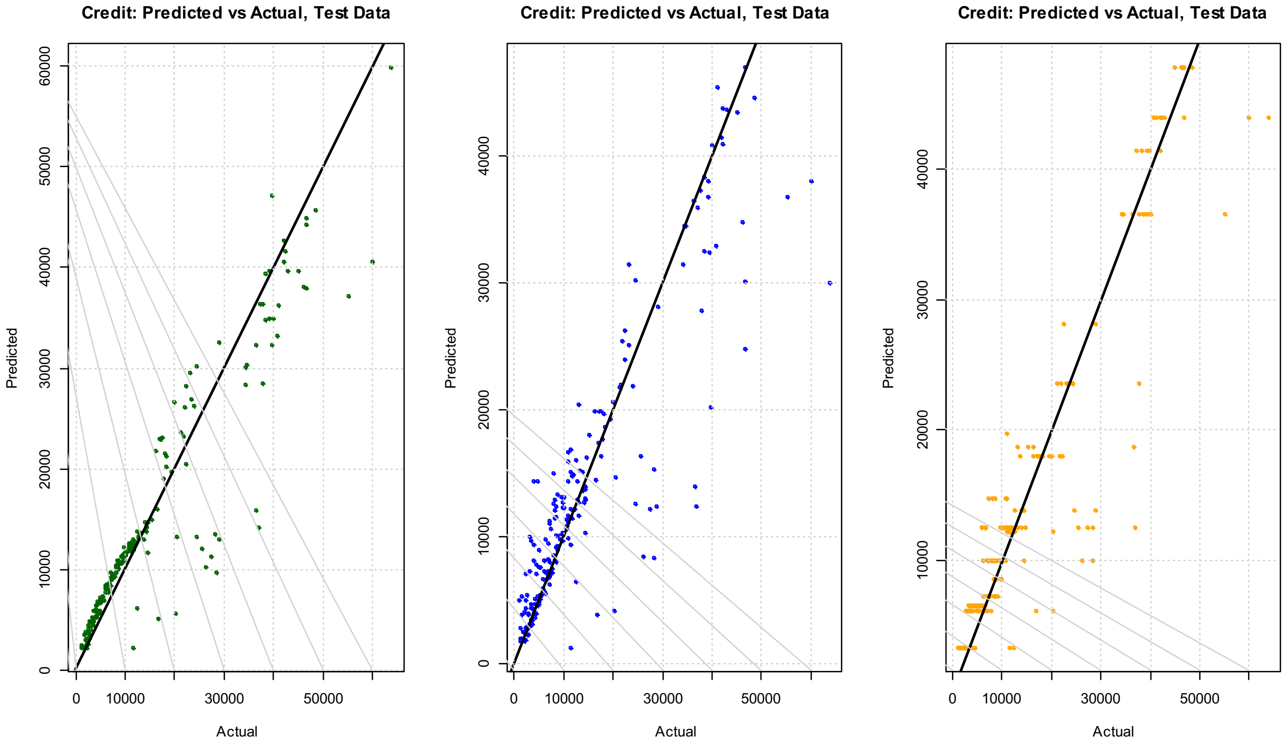

Figure 12-Visualization of 3 models

From the illustrated data in Figure 12, we can assume that the KNN model must not be the optimal because of the frequent discrepancies in the predicted data. between Linear and Regression tree we have to check the RMSEs.

Linear	KNN	Tree
4926.335	6140.336	4976.345

Table 5

The best model out of three is the **linear** model "**insu_mod_5**" which consist of all the features with their binary interactions. as we can see the two models of regression tree and linear have close RMSE and both seem to be good models.

Final

At the end we fit the candid model on the test data to calculate the RMSE.

the RSME of final candid model is 5041.705

Conclusion

This Report is created by Rmarkdown and aims to illustrate the analysis and reasons behind the chunks of R codes of Machin learning Regression project on insurances cost prediction and statistical calculations.

I hope you a have found this Report appropriate to your use.

Thank you for your attention.

Mina Kanaani

Resources

R code, Rmarkdown, And Excel files of data is attached to this file.

Other sources:

https://www.kaggle.com

https://medium.datadriveninvestor.com/10-machine-learning-projects-on-classification-with-python-9261add2e8a7