Midterm Project

Yijing Tao yt2785

2022-03-16

## Introduction

Finding out how to predict people’s average life expectancy in certain regions is very important now. The data in this study is collected from WHO, and will focus on immunization factors, mortality factors, economic factors, social factors and other health related factors as well.  
In this study, there are in total 21 variables and one response.

Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

## Visualization

In this project, I used the function featurePlot() in caret to visualize the data. Since within the 20 predictors, only “Status” is a binary pridictor, I excluded it when making the featurePlot.  
When observing the 19 plots I have got, I found that generally we can consider that having higher “Thiness 1-19 years”, “Thiness 5-9 years” and “HIV/AIDS” will lead to a lower life expectancy, while having a higher “Income.composition.of.resources”, “School”, “Polio”, “Total.expenditure”, “Diphtheria”, “GDP” will lead to a higher life expectancy.  
Then I made a density plot to show the relationship between Status and Life expectency. From the plot we can generally find that people living in developed regions seems to have longer life.  
The trends listed above are all seemed reasonable in the common sense. But what make me feel strange are that from the plots we can generally find that having a higher “BMI” and “Alcohol” will lead to a higher life expectancy, while the value of “Hepatitis.B” and “Measles” seems have no relation to the life expectancy.

## Model

After having the dataset, I included all of the variables expected of “Country” since the country name should not be an important factor which will affect the life expectancy.  
In this study, we can get a data frame which includes 1649 rows and 21 columns after omitting all of the NA values and the variable “Country”. Then I randomly extracted 70% of the data to be train data and the 30% rest to be test data.

To find out the relationship between different predictors and life expectancy, I decided first to find out the best fitted model. So I built KNN, linear regression, ridge, lasso, elastic net, PCR, PLS, GAM, MARS in total 9 models.

## weakness of different models

The first model I used to fit is KNN model. The weakness of KNN is  
1. High computational complexity; high spatial complexity.  
2. Low prediction accuracy for rare categories when the sample is not balanced  
3. Poor interpretability, cannot give rules like decision trees.  
By calculating the test error using the test data set, we can finde that the test error of the KNN model is extremely high. Therefore, I think the KNN model is not flexible enough to capture the underlying truth.

Then I used linear regression model, whose weakness is being difficult to interpret the correlation coefficient if the features are highly correlated.

The weakness of ridge, lasso and elastic net model is introducing a small amount of bias into the model, but greatly reduces the variance.

Although MARS has the weakness of requiring strict assumptions and the need to deal with outliers, MARS is not only highly adaptive compared to other methods, but also has a higher accuracy for model prediction. In the multidimensional case, due to the expansion of the sample space, how to divide the space becomes a crucial issue. MARS is a regression method with high generalization ability specifically for high-dimensional data. This regression method uses the tensor product of the spline function as the basis function, and the determination of the basis function (the number of tensor variables and the partition point of the variables) and the number of basis functions are done automatically by the data, without manual selection.

## tuning parameters

Both Ridge and lasso has 1 tuning parameter , and elastic net model has 2 tuning parameters and (0-1). Based on the rule, I tried different area of lambda and finally decided to set -2 < < 5 as the area of , and get the best tuning parameters , , , .

In the PCR and PLS model, the tuning parameter is the number of predictors included in the final model (with the smallest RMSE). In the PCR model, all of the 20 predictors are considered to be included, while in the PLS model, the model including only 18 predictors has the smallest RMSE.

In MARS model, after trying several times, I decided to take degree = 1-5, nprune = 10-30 to be the area of the tuning parameters. After training the model with caret package, it is reported that degree = 2 and nprune = 21 is the best tuning parameters that will lead to a model with the smallest RMSE.

## comparison

To find out the best fitting model, I compared their goodness of fit by comparing the RMSE. The result of the comparison through cross validation is below.

From the comparison of both cross validation and the test error, we can find that the MARS model is the best model to our data set.

## pdp plot get the important predictors

By making the pdp plot of MARS model, we can find that having a higher **“Adult Mortality”, “Under five death”, HIV/AIDS” and “Thinness 5-9 year”** will lead to an observable decrease in life expectancy, while having a higher **“Diphtheria”** will lead to an observable increase in life expectancy. So these variables above play important roles in predicting the response.  
The other variables such as “Year”, “Alcohol” etc. seems have no observable relationship with life expectancy in this study.

The main questions we want to answer by analysing this dataset is: 1. Does various predicting factors which has been chosen initially really affect the Life expectancy? What are the predicting variables actually affecting the life expectancy?  
Not all of them. By making the pdp plot of MARS model, we con find that having a higher “Adult Mortality”, “Under five death”, HIV/AIDS” and “Thinness 5-9 year” will lead to an observable decrease in life expectancy, while having a higher “Diphtheria” will lead to an observable increase in life expectancy. So these variables play important roles in predicting the response.  
The other variables such as “Year”, “Alcohol” etc. seems have no observable relationship with life expectancy.

1. Should a country having a lower life expectancy value(<65) increase its healthcare expenditure in order to improve its average lifespan?  
   No, from the model, healthcare expenditure is not strongly related to the life expectancy.