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 of this data set 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 of the Data Set

After having the data set, I included all of the variables except “Country”, “Year” and “Alcohol” in the training and testing data set since the country name and repeated “Year” should not be an important factor which will affect the life expectancy, and the personal consumption of alcohol is too small and too similar in this data set that I think it might not properly show the true alcohol consumption.  
In this project, I used the function featurePlot() in caret to visualize the data. Since within the 18 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” might lead to a lower life expectancy, while having a higher “Income.composition.of.resources”, “School”, “Polio”, “Total.expenditure”, “Diphtheria”, “GDP” might 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 is that from the plots we can generally find that having a higher “BMI” will lead to a higher life expectancy, while the value of “Hepatitis.B” and “Measles” seems have no relation to the life expectancy.

## Models Training

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. In all of the model, I conducted 10-fold cross validation method to get a better model.

## weakness of different models

### KNN

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.  
In KNN model, the tuning parameter is “k”, after training with the area (1,20), we can learn that the best tuning paremeter k is 12. Then I input the training data set and arrange y = life expectancy, x = 19 variables.  
By calculating the test error using the test data set, we can find 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.

### LM

Then I used linear regression model, whose weakness is being difficult to interpret the correlation coefficient if the features are highly correlated.  
Similar to KNN model, there is no any tuning parameters, so I simply input the training data set and arrange y and x.  
By calculating the test error using the test data set, we can find that the test error of the linear regression model is also high. Therefore, I think this model is also not flexible enough to capture the underlying truth.

### Ridge, Lasso and Elastic Net

The weakness of ridge, lasso and elastic net model is introducing a small amount of bias into the model, but greatly reduces the variance.  
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 , , , .  
By calculating the test error using the test data set, we can find that the test errors of these 3 models are small. Therefore, I think these 3 models are also flexible enough to capture the underlying truth.

###PCR and PLS 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 18 predictors are considered to be included, while in the PLS model, the model including only 17 predictors has the smallest RMSE.  
By calculating the test error using the test data set, we can find that the test errors of PCR and PLS model are small. Therefore, I think these models are flexible enough to capture the underlying truth.

### GAM

The weakness of GAM model is lack of parametric functional form makes it difficult to score the new data directly. The tuning parameter of GAM model is whether the “select”. If it is “TRUE” then GAM can add an additional penalty variable to each semester so that it can be scored as zero. This means that the smoothing parameter estimate is part of the fit and can be completely removed from the terms in the model. If the corresponding smoothing parameter estimate is zero, then the additional penalty has no effect. In this project, the “select” is “FALSE”.  
By calculating the test error using the test data set, we can find that the test error of GAM model is small. Therefore, I think this model is flexible enough to capture the underlying truth.

### MARS

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.  
In MARS model, after trying several times, I decided to take degree = 1-5, nprune = 10-29 to be the area of the tuning parameters. After training the model with caret package, it is reported that degree = 2 and nprune = 26 is the best tuning parameters that will lead to a model with the smallest RMSE.  
By calculating the test error using the test data set, we can find that the test error of MARS model is small. Therefore, I think this model is flexible enough to capture the underlying truth.

## 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.

## Important Predictors

By making the vip importance plot of MARS model, we can find that changes in **“Income composition of resources”, “Adult Mortality Rates”, “HIV/AIDS”, “Thinness 5-9 years”, “Diphtheria”, “Infant death”, “Status(developing)”** and **“BMI”** will lead to an observable change in life expectancy. So these variables above play important roles in predicting the response.  
The other variables seems have no observable relationship with life expectancy in this study.

From the slope of pdp partial plots and the coefficient, we can find than among the 8 important predictors, a higher “Adult Mortality Rates”, “Infant death”, “HIV/AIDS”, and being a “developing” country will all lead to a lower life expectancy (the coefficient of “StatusDeveloping \* h(Adult.Mortality-118)” is smaller than 0).  
“Income composition of resources”(ICR), “BMI”, “Diphtheria” and “Thinness 5-9 years”(T5.9Y) don’t have a monotonous influence on life expectancy – when the predictor is smaller than 0.3 and larger than 0.8, the increase of “ICR” will lead to a lower life expectancy while when “ICR” is large than 0.3 and smaller than 0.8, the increase of it will lead to a higher life expectancy. When “BMI” is smaller than 45, the increase of “BMI” will lead to a higher life expectancy while when “BMI” is large than 45, the increase of it will lead to a lower life expectancy. When “Diphtheria” is smaller than 60, the increase of “Diphtheria” will not lead to any change in life expectancy while when “Diphtheria” is large than 60, the increase of it will lead to a higher life expectancy. When “T5.9Y” is smaller than 5, the increase of “T5.9Y” will lead to a lower life expectancy while when “T5.9Y” is large than 5, the increase of it will only lead to a slightly increase in life expectancy.

## Discussion

Compared to the interpretation I have made in the visualization, some of the variables I thought would be important to the result, such as “Polio” and “GDP”, was not included in the final model. However, in my own point of view, this might because the sample size is not large enough, or they are not as important as the 8 variables which are included in the final model, so I actually don’t think this means that they are completely not related to the life expectancy.