Chu_Yiwei_yc878_finalproject

Yiwei Chu

12/2/2019

Introduction

- The aim of the project

  The project focuses on alcohol use disorders(AUD). I choose several related variables as predicators to predict the percentage of population with alocohol use disorders all over the world.

  Alcohol use disorders have aroused national concern these days. As people with alcohol use disorders are younger, the accidents caused by alcohol use disorder becomes more and more. Therefore, predicting the percentage of population with alcohol use disorders, examining what kind of variables may influence the alcohol use disorders, and thus controlling the alcohol use disorders all over the world are important and have great significance.

  The goal of the project is to find as many variables as I can and predict the percentage of population with alcohol use disorders precisely. Through calculating the MSE, I can know how precise the predication is and whether I find appropriate variables as predicators or not.

- The efforts I have made for the project

  In the first step, I found my huge interest in and great concern on the use of alcohol disorder all over the world. I searched a lot of databases to get appropriate data and decided the topic of alcohol use disorder. Compared to the alcohol consumption I

wanted to do at first, the topic of alcohol use disordesr has policy significance and provides more guidance and policy suggestions for the control of alcohol consumption.

The second thing I did was looking for data and cleaning up the data. I prefer to explore the influencing factors of alcohol use disorders from the perspective of social science, therefore, I chose Drug Use Disorder, Age Dependency Ratio, GDP, GNI, Labor Force, Primary School Enrollment, Refugee Population, Unemployment, Population and Sex Ratio as variables from two data sources: Our World in Data and Worldbank. After deciding what data and variables I want to use, I cleaned up the data because the data is dirty and has a lot of unnecessary information. Also, I made visualizations for these variables to see the distributions of these variables. The next several parts will cover more details of cleaning up the data.

The third step I did was cross-validation. I created ten folds.

The fourth step I did was running three different models to predict the percentage of population with alcohol use disorders: Linear Regression Model, K-Nearest Neighbors Model, and Random Forest Model. The methods I use is supervised learning (Regression).

The fifth step I did was to test the accuracy of the three models, and I basically calculated RMSE for testing the error of every model and Rsquared for examining which model is the most appropriate. After choosing the most fit model, I calculated MSE for that model to see how precise the model is.

In step six, I checked variable importance to see what predicators have most influence on percentage of populations with alcohol use disorders.

By doing these six steps, the project was able to arrive at a complete conclusion and make accurate predictions.

Problem Statement and Background

- Problem statement and background

  It's not easy to realize that alcohol use can cause problems. Drinking is socially acceptable in most places and is often used as a social lubricant. Even though alcohol use disorder and other substance use problems are considered diseases like any other, many people choose to interpret alcohol use disorder from a moral perspective. AUD is caused by complex interactions between genes and the environment, which are closely related to other health problems. For example, people with alcohol use disorders are more likely to have other illnesses such as cancer and mental illness such as depression and mania. Although genes play an important role, exposure to certain life events and situations significantly increases a person's vulnerability to using alcohol for comfort and reward.

  In this project, I hope to research percentage of population with alcohol use disorders because it has been a serious social issue and many friends of mine are suffering from the disorder. I collected data from Worldbank and our world in data. After cleaning it, each row is a country's name for different years from 2008 to 2017 and each column represents the variables I selected. It has 1067 rows of observations and 15 columns. I choose Drug Use Disorder, Age Dependency Ratio, GDP, GNI, Labor Force, Primary School Enrollment, Refugee Population, Unemployment, Population and Sex Ratio as predicators.

  I hope to calculate correlation to study the relationship between attributes, the importance of predictors, and create different models(Linear Regression Model, K-Nearest Neighbors Model and Random Forest Model) to predict the percentage of population with alcohol use disorder.

- Light Literature Review

  There are not many researches using machine learning to examine and predict per-

centage of population with alcohol use disorders from the social sceince perspective. Most of the literatures are from the perspective of curing alcohol use disorders, calling for the society to pay attention to alcohol use disorders, studying the relationships between alcohol use disorders and other diseases.

Data

- Data sources

  World Bank, *Our World in Data*

- Unit of analysis

  Country_Year

- Variables of interest

  AUD(Alcohol Use Disorders), DUD(Drug Use Disorders), Age Dependency Ratio, GDP, GNI, Labor Force, Primary School Enrollment, Refugee Population, Unemployment, Population, Male Population, Female Population.

- Steps to wrangle data

  Before wrangling data, the first step I took is making clear what I need from data and having a genral view of what the data should look like. I need the data of variables of all the countries all over the world from 2008 to 2017. I need different columns to be the different variables, and each row represents country names for different years.

  The second step is to select the year and variables I need using *"select"* function. I select the year from 2008 to 2017 and select the column of country name and the column of variable I need of different data. Also, in order to select the variables and years I need, I use *"filter"* function.

  The third step I took is using the function *"gather"* to change the columns into rows.

  The fourth step is renaming the columns using *"name"* function

The fifth step is *"merge_all"*. I merge the 11 data frames into one data frame.

The sixth step is *"mutate"*. I create a new column called *"sex_ratio"* using the population of male/the population of female since the two variables, the population of male and the population of female, are highly correlated with population variable, which may lead to error when running linear regression model. However, I don't want to drop the two variables because I assume the female population and male population has great influence on the percentage of population with alcohol use disorders.

The seventh step is dropping all the NA using *"na.omit"* function. The reason why I drop all the NA is that NA in these variables doesn't have any significances. For example, the missing value for percentage of population with alcohol use disorder and drug disorder of some countries, it is a mistake to impute it or fill it with mean or median. What I found is that most of the missing value of all the variables is concentrated in a few countries, most of which are small countries.

Analysis

- The methods/tools I explored

  I firstly use data wrangling to clean the data, which we have mentioned before. I secondly use data visualization for the distribution of variables and the correlation of variables. Thirdly, I use supervised learning (regression), running linear regression model, K-Nearest model and random forest model to predict the percentage of population with alcohol use disorders. Fourthly, I calculated Rsquared and RMSE to compare the model with best performance. Fifthly, with the best performance model, random forest model, I calcualted its MSE to see how precise its prediction is.

- Detailed entire analysis

  When we want to explore the relationships among variables and predict the outcome in an efficient way, we use regression supervised learning when the outcome we predict

is numeric.

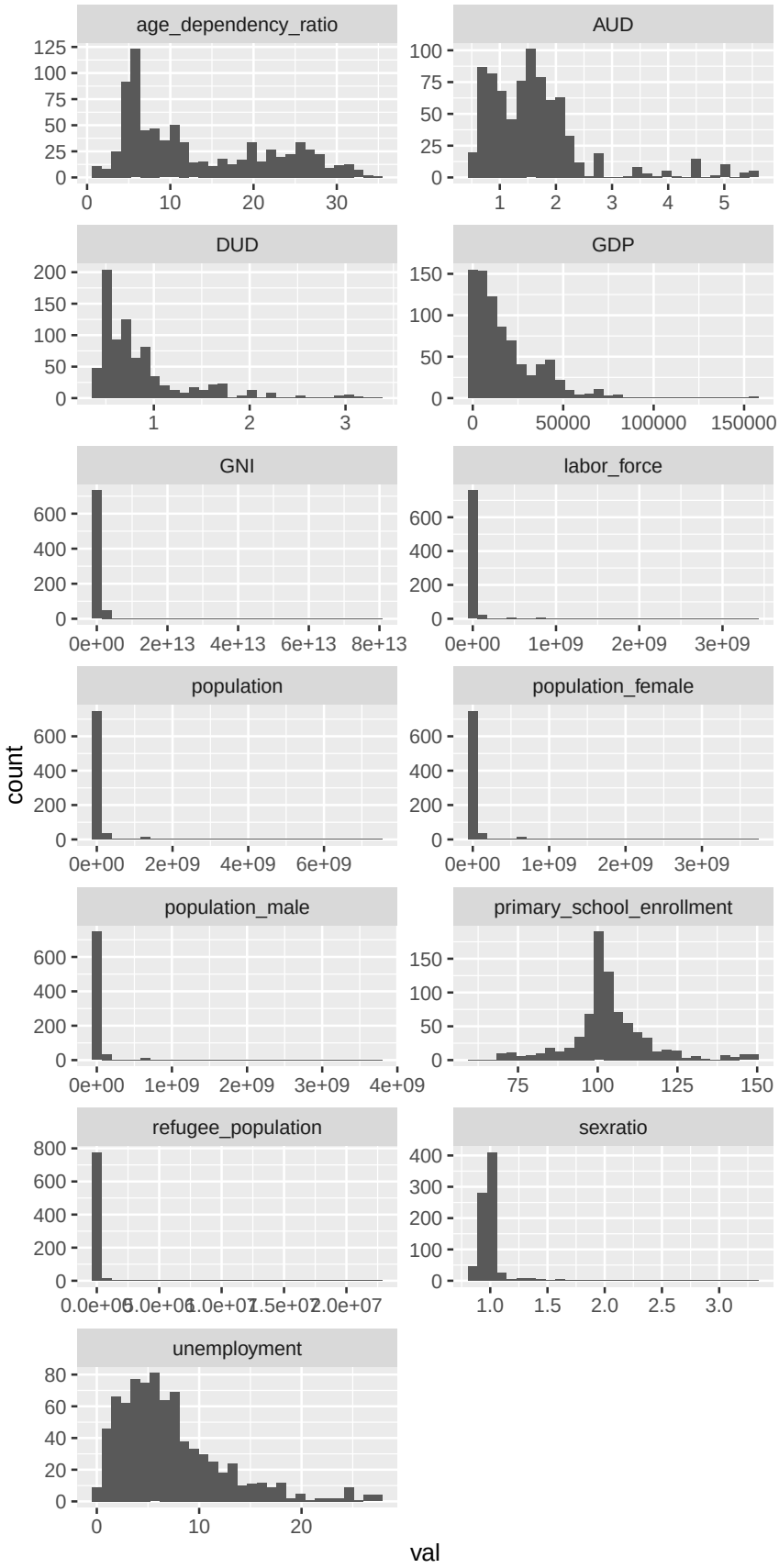
Linear regression model is a model that studies the dependence of one variable on another or other variables. In this project, it has the worst performance with largest error and smallest Rsquared.

For K-Nearest model, it means that each sample can be represented by its nearest k neighbors. The core idea of K-Nearest model is that if the samples' k nearest neighbors belong to a certain category, then the sample also belongs to this category and has the characteristics of samples in this category.

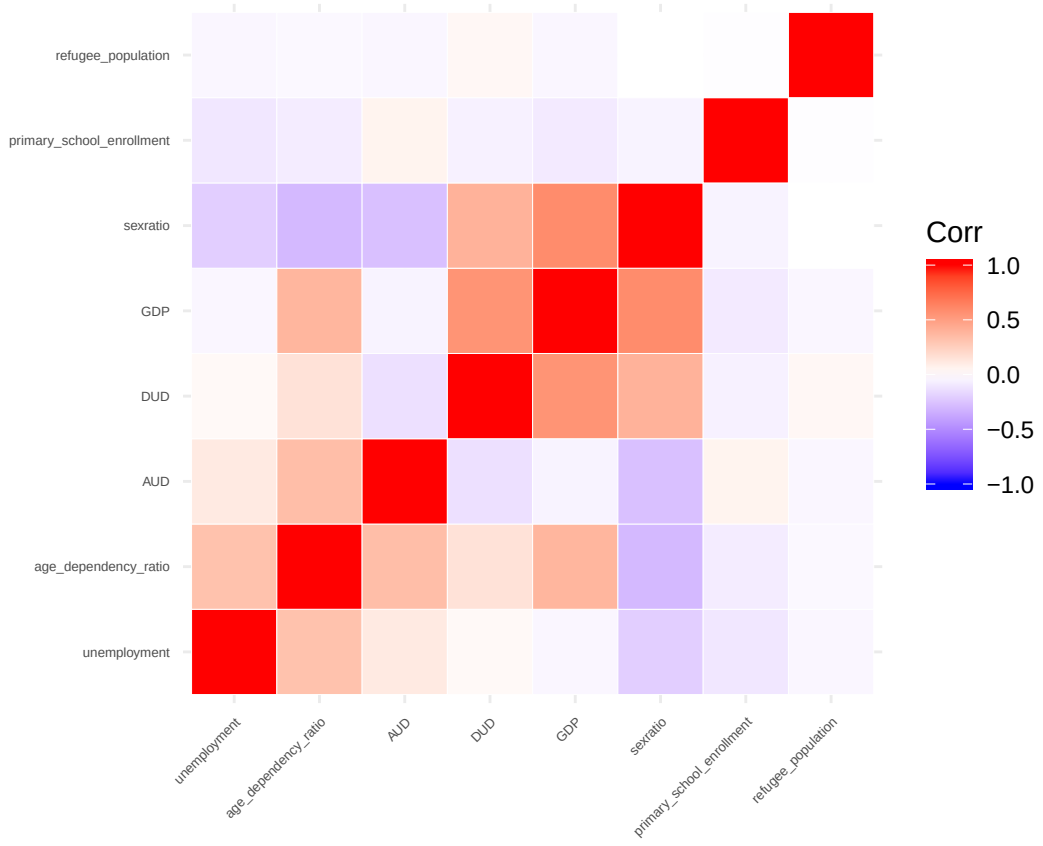
For random forest model, it is an excellent model.It has high efficiency for the the regression and classification of multi-dimensional feature data sets, and can also make the selection of feature importance. The operation efficiency and accuracy are high, and the implementation is relatively simple. In this project, random forest model has the best performance with least error, highest R-squared and MSE of 0.1469433. These all mean that random forest model predicts precisely in this project.

Before running these models, I use K-fold cross-validation to test the accuracy of algorithms. I divide the data set into ten folds, among which 9 parts are used as training data and 1 part is used as test data in turn to carry out experiments. The corresponding correct rate is obtained for each experiment. The average of the accuracy (or error rate) of the results of 10 times is used as the estimation of the accuracy of the algorithm. In general, multiple 10-fold cross validation (such as 10-fold cross validation for 10 times) is needed to obtain the mean value, which is used as the estimation of the accuracy of the algorithm.

I at first see the distribution of the variables. Most of the data are right-skewed, and only the distribution of primary school enrollment is close to the normal distribution.


age_dependency_ratio
AUD
DUD
GDP
GNI
labor_force
population
population_female
population_male
primary_school_enrollment
refugee_population
sexratio
unemployment
count
val

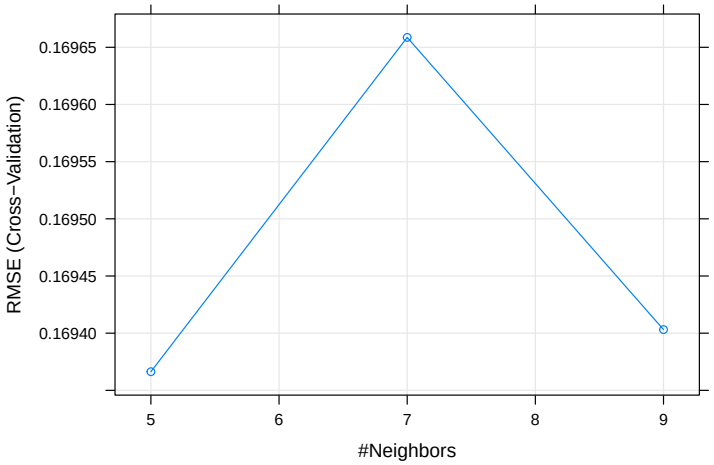
I then visualize the correlation among selected predicators. The darker color means that the two variables have the more stronger relationships. For example, the sex ratio has stronger realtionship with the country's GDP and drug use disorder and has weaker correlation with age dependency ratio.

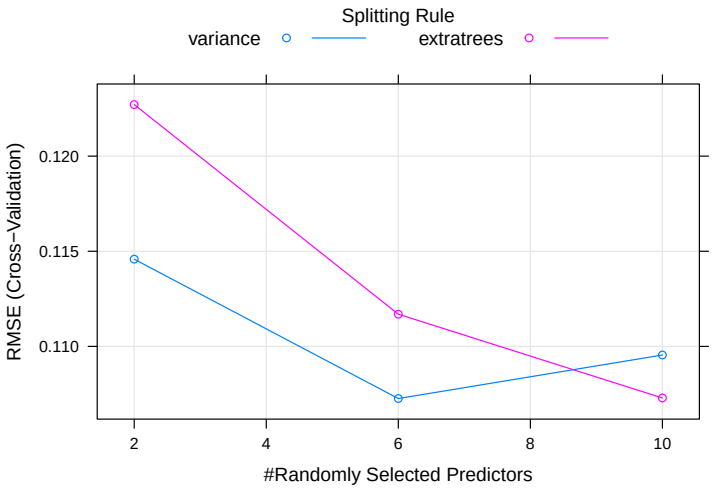


Results

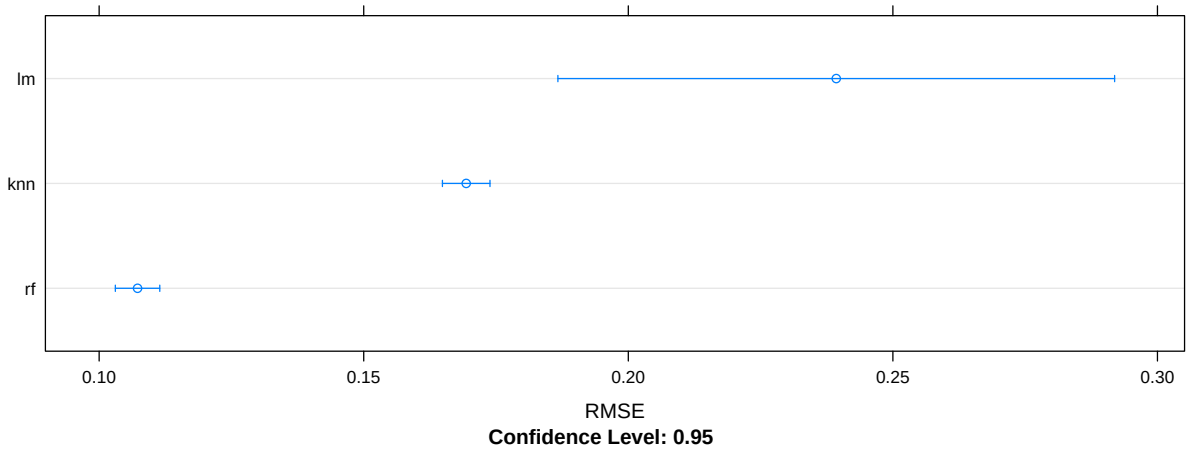
Among the three models, random forest model has the best performance.

The following two visulaization for K-Nearest Model and random forest model. For K-Nearest model, we can see that when k equals to 5, it has the smallest RMSE, while k equals to 7, it has the largest RMSE, which means that when the model has five neighbors, the model has the best performance, and when the model has seven neighbors, the model has the worst performance.

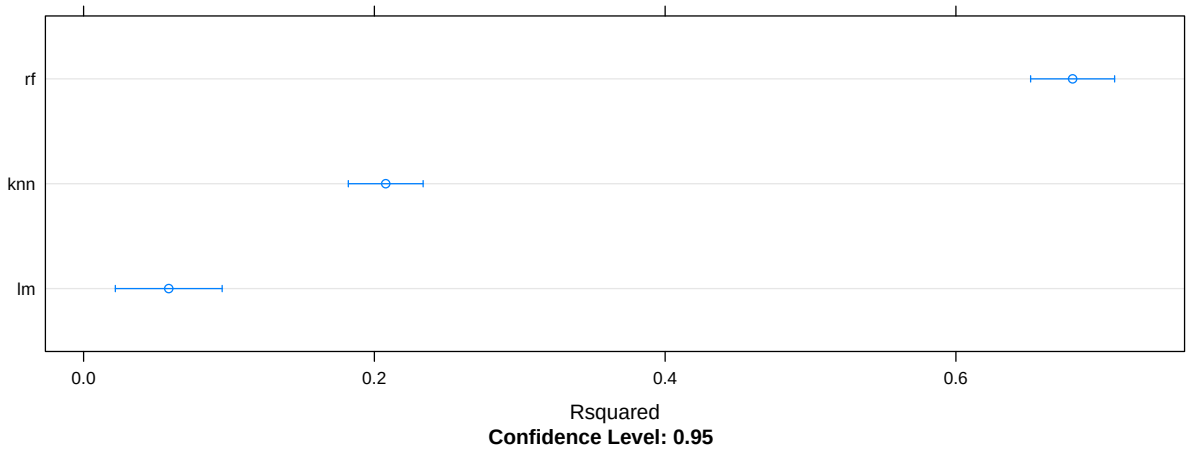




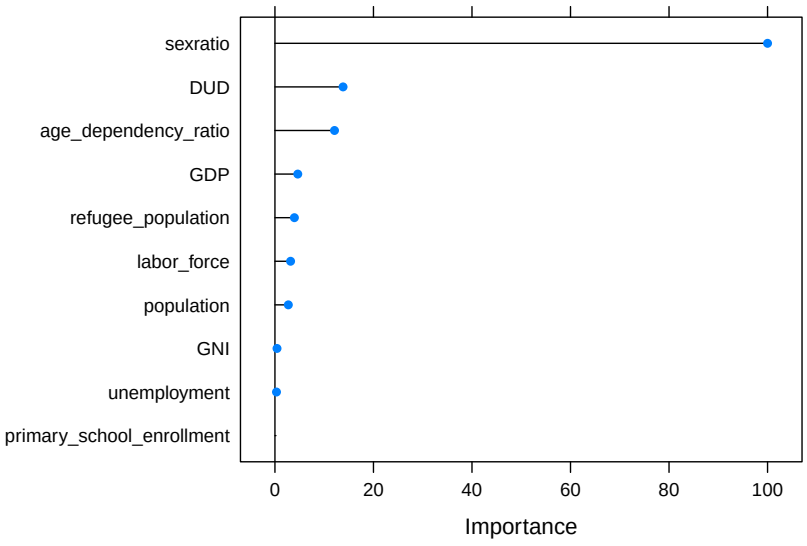
I calculate the error of each model. The linear regression model has the biggest RMSE, meaning that it is the model with worst performance.



Random forest model has the biggest Rsquare, meaning that it performs the best. Therefore, we will choose Random forest model. After being sure that random forest model is the best fit model, I calculate its MSE to see how precise its prediction is. The MSE is only 0.146933, meaing that it predicts pretty precisely.



For the variable importance, the first three most important variables are: sexratio, drug use disorders, and age dependency ratio. It means that the three have larger influence on the percentage of population with alcohol use disorders.



Discussion

- "Success" of the project

  This is a successful project because it predicts the percentage of population with alcohol use disorders with a very small MSE(0.1469433). This small MSE means that the Random forest model predicts precisely. I choose the variables successfully. After examining the variable importance, the variable of sex ratio is the most important which has nearly 100 importance. Moreover, after comparing the three models, Isuccessfully compared several models to find a model with better performance (Random forest model).

- Tools/methods I considered but did not use

  I was thinking of to change the skewness of the variables since most of the variables are right-skewed.

  I was planning to use boxplot to analyze whether there are an unreasonable outliers in the data. In that case, I can clean up the data for further data cleansing.

- Expanding the analysis

  I will use boxplot to analyze whether there are an unreasonable outliers in the data and clean up the data further.

  I will use variables with higher correlation or importance to create the model and comparethe performance of the model after dimensionality reduction and the model before dimensionality reduction.

  I will repeat k cross validation to increase the accuracy of the model.