POTABLE WATER PREDICTION

Performed by:

Aanchal Gupta

CONTENTS

1. Introduction
2. About the Data set
3. Data
    a. Read data
    b. Clean names and Convert the target Var to factor
    c. Summary
4. Missing data
    a. Missing Data Count
    b. Missing data rate
    c. Missing data rate VS Target Variable
    d. Replace Missing data with mean
5. Outliers
    a. Box Plot and Jitter
    b. Histogram
6. Correlation
    a. Corrplot
    b. ggpairs
7. Caret Random Forest Model
    a. Data Split, 80% Training Data
    b. Best mtry
    c. Train Random Forest Model
    d. Important variables
    e. Prediction
8. Final note

Introduction

We all know the importance of water, and I do not find any reason to repeat what everybody has said about the benefits of drinking water. (water equals life itself). We all certainly know that there is a drinking water crisis that appears in many countries to the extent that many countries are buying drinking water and transport it through water tankers. The crisis will only worsen in the future.

Although it would take me a long time to create an accurate model, I decided to build one for this dataset, I always try to be very careful in selecting and testing the factors that affect the results and the accuracy generated by the model.

About the Data set

There are ten variables in this dataset; nine are numeric type, and one a factor type (0 and 1) 0 means the water sample is not fit to drink.

The data set contains water quality metrics for 3276 different water bodies.

1. **pH value: pH** is an important parameter in evaluating the acid–base balance of water.
2. **Hardness:** Hardness is mainly caused by calcium and magnesium salts.
3. **Solids (Total dissolved solids - TDS)**: Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc.
4. **Chloramines:** Chlorine and chloramine are the major disinfectants used in public water systems.
5. **Sulfate:** Sulfates are naturally occurring substances that are found in minerals, soil, and rocks.

6. **Conductivity:** Pure water is not a good conductor of electric current rather’s a good insulator. Increase in ions concentration enhances the electrical conductivity of water.
7. **Organic_carbon:** Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources.
8. **Trihalomethanes:** THMs are chemicals which may be found in water treated with chlorine.
9. **Turbidity:** The turbidity of water depends on the quantity of solid matter present in the suspended state.
10. **Potability:** Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

DATA

In the dataset, there are 3276 rows and 10 columns. No doubt, there must be more parameters than what we have in this dataset, such as Arsenic, Cyanide, Lead, Zinc, Aluminium, etc. The potability (target) variable is an integer type and should be converted to Factor type. Also, we might have normalized the data, but since I aimed to use Random Forest classification (decision tree) I do not see a reason to do normalization.

Missing Values

Handling missing values is one of the common tasks in data analysis. I have two ways to deal with them.

- Remove them: I will lose a significant number of records, and this may harm the model, or
- Replace them with the mean.

After many experiments, I decided to choose the second option.

Outliers

I have two options:

- Remove them: Not sure about the benefit from this choice, or
- Keep them: If they have no bad leverages.

So, for understanding what needs to be done in this regard, I used boxplots and histogram. From these I could conclude that nothing is critical, and the outliers, in my opinion, will not have heavy leverages as the total observations are reasonably high.

Correlation

To check, used the correlation matrix to find the correlation coefficients of the independent variables in this data set. I wanted to see if we have multicollinearity problem. In this case, I decided to use glm() function from glm package.

I concluded, No multicollinearity. The predictors are not strongly correlated, so we may keep them all as they are.

Caret Random Forest Model

To use the random forest algorithm, I firstly split the data according to the 80:20 ratio. So, 80% of the data was used as Training Data and the remaining 20% as the Test Data.

The next step was to choose the best mtry. The mtry parameter is the number of variables available for splitting at each tree node. In the randomForest package used in R, for classification models, the default is the square root of the number of predictor variables (rounded down). Here, we chose the mtry value as the one with the least value OOB error, out of the different possible values. (mtry=4)

After training the random forest model, we found-

- sulphate,
- ph,
- hardness,
- solid and
- chloramines

were the top five variables, according to this dataset.

Prediction

Finally, to check the performance of the fitted model, I obtained-

- Confusion Matrix
- Accuracy
- 95% Confidence interval
- Specificity
- Sensitivity

The Accuracy concluded that it is a good result. 95% confidence interval was also Fair. The Sensitivity was not high and Specificity was also acceptable, depending upon the aim of our study.

Final note

I think we need more variables to work on, and maybe we need some features engineering strategy to get better results. I worked with many methods but couldn’t improve the results more than that, but the important thing is that I refreshed my knowledge about hyper-parameter settings.

Finally, I should say that I have tried many algorithms and many methods using my local machine, like xgboost but I couldn’t get better results either, and I believe this dataset can’t give us better results for Accuracy, sensitivity, specificity and other important indicators, unless we add important features as Arsenic, Cyanide, Lead, Zinc, Alkalinity, Aluminium, Boron, etc.