final_project

2021/11/22

Machine Learning Final Project: asthma prediction

Akeem Yusuf (11165362), Connor Burbridge (11162928), Jinhao Zhong (11204178), Olorunifemi Aina (11298221), Qi Zhao (11302962)

Abstract

The data analyze study aim to identify the data and find out the fit model to do the prediction. In this report, our team is going to propose different way to select the important data from all the potential features, then we will use various prediction models to find out what are the features are significant to asthma.

1. Introduction

Asthma is a lung disease which your airways narrow and swell and may produce extra mucus. This can make people difficult to breath and cause coughing, wheezing and short breath. We do not know all the reasons that can cause asthma, thus, in this project, we are going to analyze the relationship between gender, BMI, age, smoking status, forced expiratory volume and 500 genotyping. Because there are large amount of variables, our team is going to use Lasso, t- test and random forest to do the feature selection to find out the significant variables, then we will LDA, QDA and Elastic Net on the fitting models to do the prediction. In order to get the best asthma prediction, our team would compare the root-mean-square (RMSE) and choose the least RMSE as our result.

2. Data

2.1 Data Description

There are 506 variables and 112167 observations in the test set, and 509 variables and 112151 observations in the training set. There are three outcome variables in the data set, including Asthma status, COPD status and Cancer status. As a result, the relevant variables are referenced as potential features. The potential features and outcome variables are shown as follows:

outcome variables(0 = yes, 1 = no): Asthma Status, COPD Status, Cancer Status

Potential Features:
Sex: Female = 0, Male = 1
BMI: body mass index
Age
Smoking Status: Prefer not to answer = -3, Never = 0; Previous = 1, Current=2
FEV1Z: Score for FEV1 (forced expiratory volume)
SNP1-500: top 500 genotyping variants, in which 0, 1, 2 represent different SNP

2.2 Summary statistics of outcome variables and potential features

mean, median and variance

Below we can see the mean, median and variance values of BMI, age and FEVIZ, and we can see the structure of our data set in histgram. BMI:

```
## [1] 27.30576

## [1] 26.5963

## [1] 22.18281
```

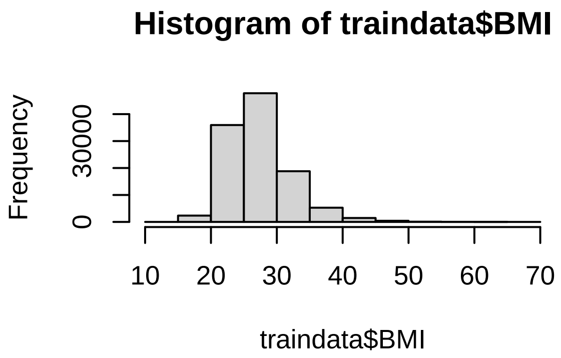
Age:

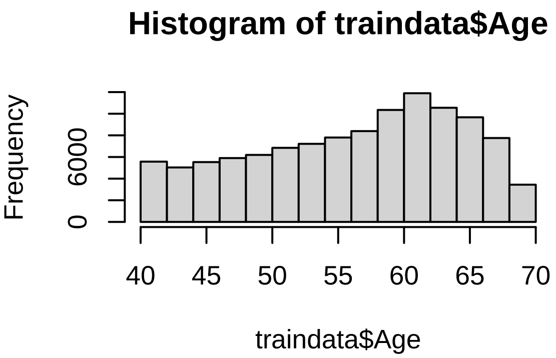
```
## [1] 56.65118

## [1] 58

## [1] 62.72659
```

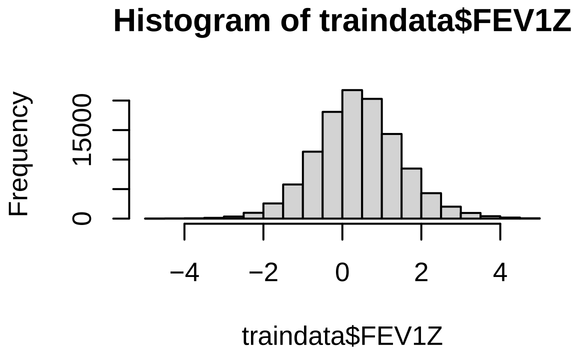
FEVIZ:

```
## [1] 0.4067527

## [1] 0.386

## [1] 1.18259
```

We can see that most of the people has BMI between 25 to 30, and second most is 20 to 25, and the largest group of age is between 60 and 65. Also, most of the people has the similar FEVIZ.







From the data below, we can find out there are about 15% more female than male in our data set. Secondly, most of the people never smoked before, also, there are very low percentage people in database has asthma, it is only about 13.47 percent.

percentage of male:

```
## [1] 42.636
```

percentage of female:

```
## [1] 57.364
```

pencentage of never smoke:

```
## [1] 54.652
```

percentage of previous somking:

```
## [1] 36.766
```

percentage of currently smoking:

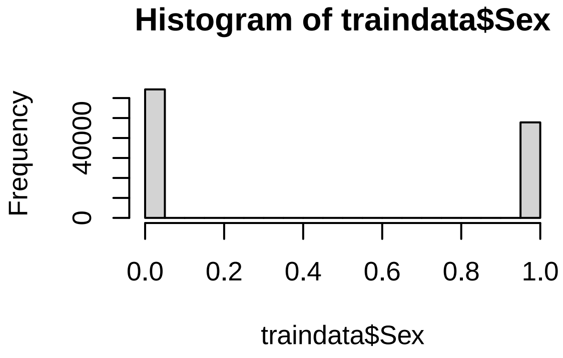
```
## [1] 8.582
```

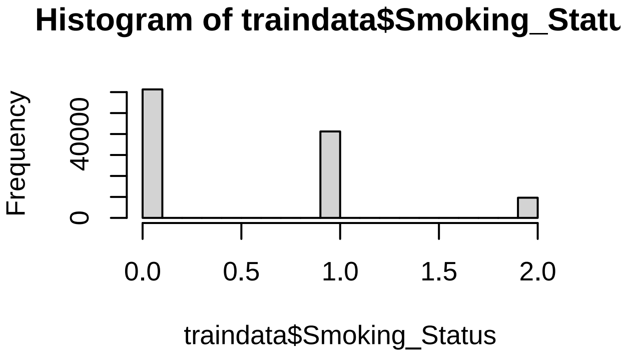
percentage of having asthma:

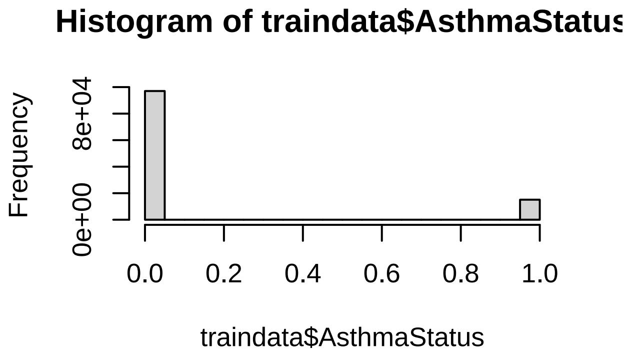
```
## [1] 13.47
```

percentage of not having asthma:

```
## [1] 86.53
```







Methodology and Research

3.1 Identifying features

In this section, we are going to use different method for feature selection.

t-test t-test can be used to test whether two group data is statistically significant different. The null hypothesis (H0) is that the difference between the two-group means is zero. We compare the p-value get from t-test with the 0.05. If p-value<0.05, reject the null hypothesis. If the values of a predictor responding to healthy samples (Asthma=0) are significantly different from the values of the predictor responding to the disease samples (Asthma=1), the predictor is meaningful to predict Asthma.

Lasso

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{1}$$

is the the standard linear model. Where Y is a response. X1, X2, X3, . . . are predictors. ϵ is error.

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \tag{2}$$

is residual sum of squares (RSS).

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \tag{3}$$

is the constraint for Lasso. Lasso estimates β_0, β_1, . . . , β_p using the values that minimize $RSS + \lambda\sum_{j=1}^{p}|\beta_j|$ if the coefficient of Xi is zero, Xi is meaningless to predict response.

3.2. Method selection

In this section, we will introduce the prediction models that we use.

LDA

LDA estimate probabilities by using Bayes' Theorem. LDA predicts response by calculating the following formula:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{x-\mu_k}{\sigma})^2}}{\sum_{l=1}^{K}\pi_l \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{x-\mu_l}{\sigma})^2}} \tag{4}$$

Where $p_k(x)$ represent the probability of response class k when X=x. π_k is prior probability for class k. μ_k is the mean, σ is the standard deviation. By selecting the maximum $p_k(x)$, LDA classify the x to class k.

QDA

Like LDA, QDA estimate probabilities by using Bayes' Theorem too. The difference between LDA and QDA is that QDA calculate the covariance matrix for each class k, LDA not. QDA classify the x by selecting the maximum $\delta_k\left(x\right)$:

$$\delta_k\left(x\right) = -\frac{1}{2}x^T\sum_{k}^{-1}x + x^T\sum_{k}^{-1} - \frac{1}{2}\mu_k^T x^T\sum_{k}^{-1} - \frac{1}{2}log|\sum_{k}| + log\pi_k \tag{5}$$

Where Σ_k is a covariance matrix for the k^{th} class k.

Elastic Net

Elastic Net estimate the parameters in $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$ by minimizing

$$L(\beta) = \frac{\sum_{i=1}^{n}(y_i - x_i\beta)^2}{2n} + \lambda(\frac{1-\alpha}{2})\sum_{j=1}^{n}\beta_j^2 + \alpha\sum_{j=1}^{n}|\beta_j| \tag{6}$$

Where α is a parameter to tune, Elastic Net same to ridge when $\alpha = 0$, to lasso when $\alpha = 1$. Elastic Net overcome a shortage of lasso that the predictor selection can be too dependent on data.

random forest

Random forest are an ensemble learning method classification and regression, it can provide improvement by constructing decision trees and making adjustment to decorrelates the decision trees. By using random forest method, we can find out the significant predictor to asthma and reduce the variance.

3.3. Model evaluation

root-mean-square error (RMSE) When we have the fit models, our team is going to use root-mean-square error (RMSE) to find the best prediction of asthma. Root-mean-square error (RMSE) is a method for measuring the difference between values predicted by models. The differences between predicted values and observed values are called residuals or errors. It can be used on measuring accuracy to compare forecasting errors of different models. RMSE is always non-negative, and if the value is 0, it will be a perfect fit for the data. In general, the lower RMSE is the better prediction model.

3.4 Results and presentation

Finally, the model with the best result will be selected, including the parameters of the model, and the evaluation of this model. A report explaining the findings of the research will be written along with a PowerPoint presentation of the main findings.