Gender Recognition by Voice

Xinyu Zhang, Xiaochi Ge, Wenye Ouyang

12/4/2017

I. Introduction

Our SMART question is: How to use classification models to recognize gender by their voice?

The research is about how to recognize gender by voice. The dataset includes the measurement of each voice sample's auditory features.

II. Data information & Exploratory Data Analysis

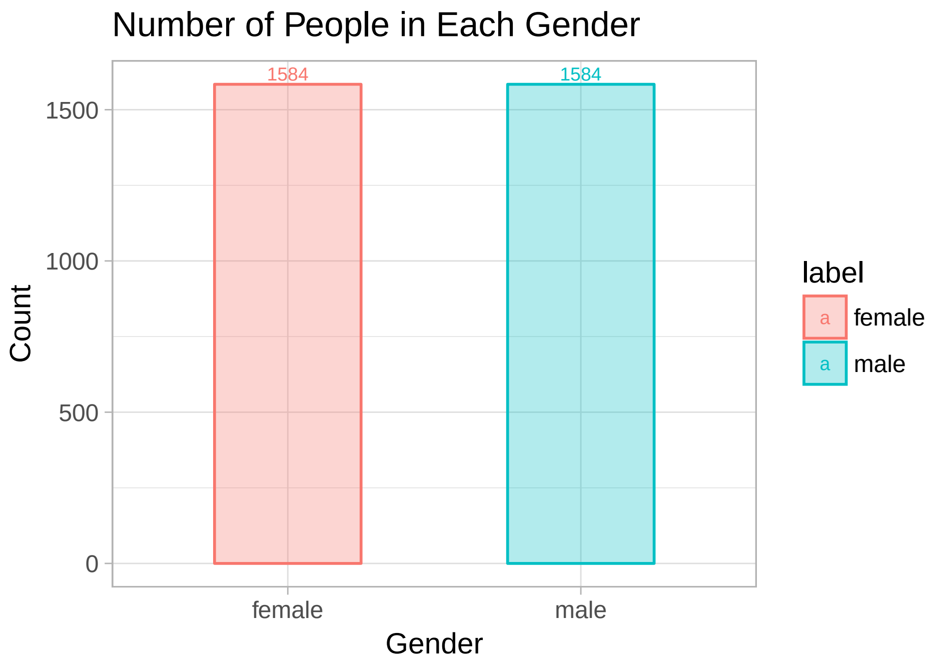
There are 21 features and one target variable `label`

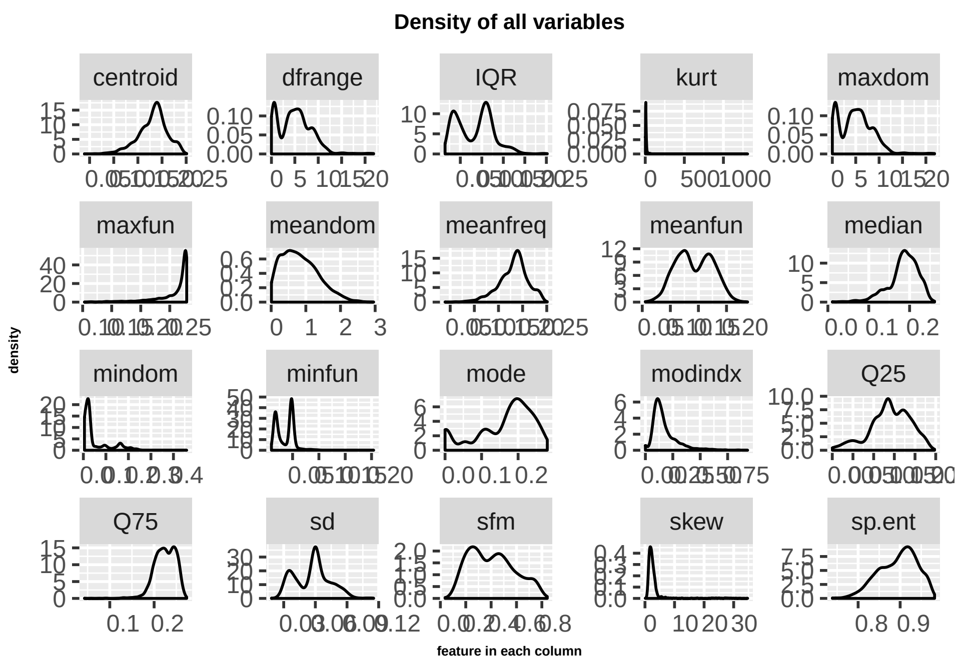
- duration: length of signal
- meanfreq: mean frequency (in kHz)
- sd: standard deviation of frequency
- median: median frequency (in kHz)
- Q25: first quantile (in kHz)
- Q75: third quantile (in kHz)
- IQR: interquantile range (in kHz)
- skew: skewness (see note in specprop description)
- kurt: kurtosis (see note in specprop description)
- sp.ent: spectral entropy
- sfm: spectral flatness
- mode: mode frequency
- centroid: frequency centroid (see specprop)
- peakf: peak frequency (frequency with highest energy)
- meanfun: average of fundamental frequency measured across acoustic signal
- minfun: minimum fundamental frequency measured across acoustic signal
- maxfun: maximum fundamental frequency measured across acoustic signal
- meandom: average of dominant frequency measured across acoustic signal
- mindom: minimum of dominant frequency measured across acoustic signal
- maxdom: maximum of dominant frequency measured across acoustic signal
- dfrange: range of dominant frequency measured across acoustic signal
- modindx: modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
- label: female and male. (This is target variable)

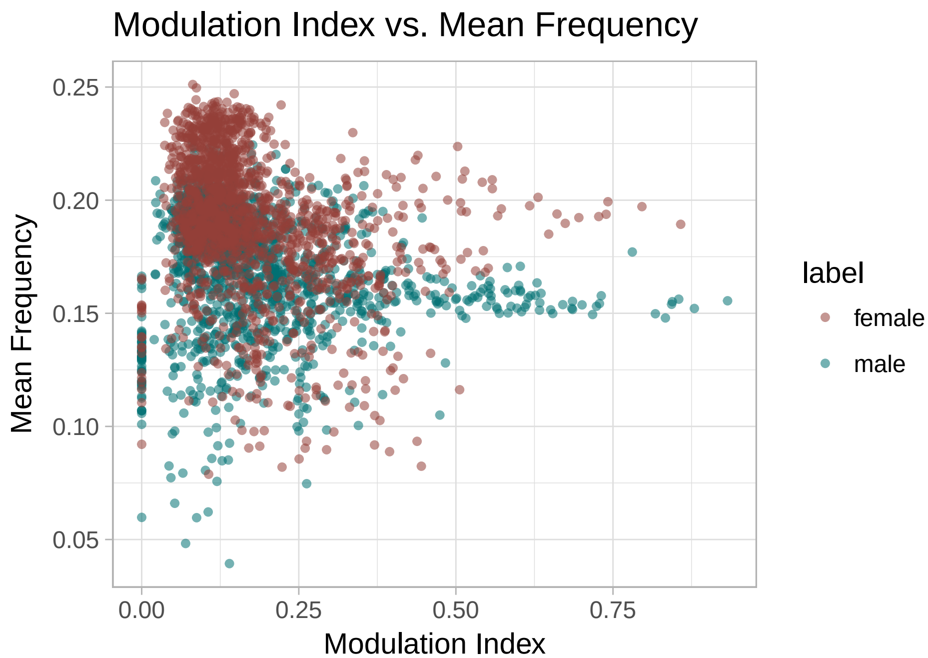
Our dataset is about the voice of gender, therefore, we conduct a question that is about how to recognize gender through voice. For a better understanding of the dataset, we do some research on people's frequency of sound.

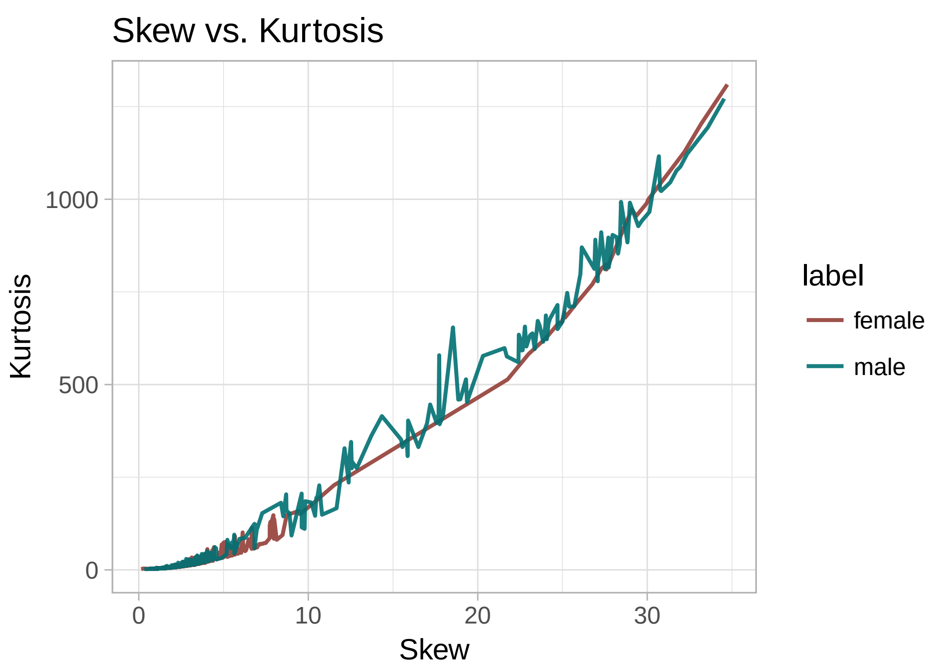
We searched online and find a voice dataset, which contains 3169 voice samples. Then, we decided to use three models (random forest, knn, and logistic) to compute the accuracy.

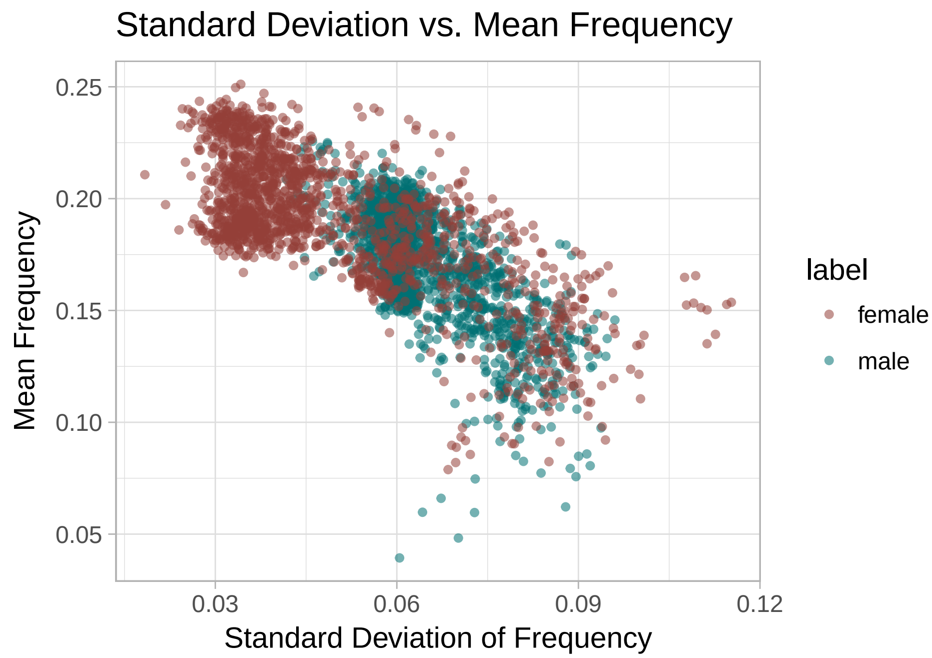
We use confusion matrix and accuracy rate to determine which regression model is the highest.

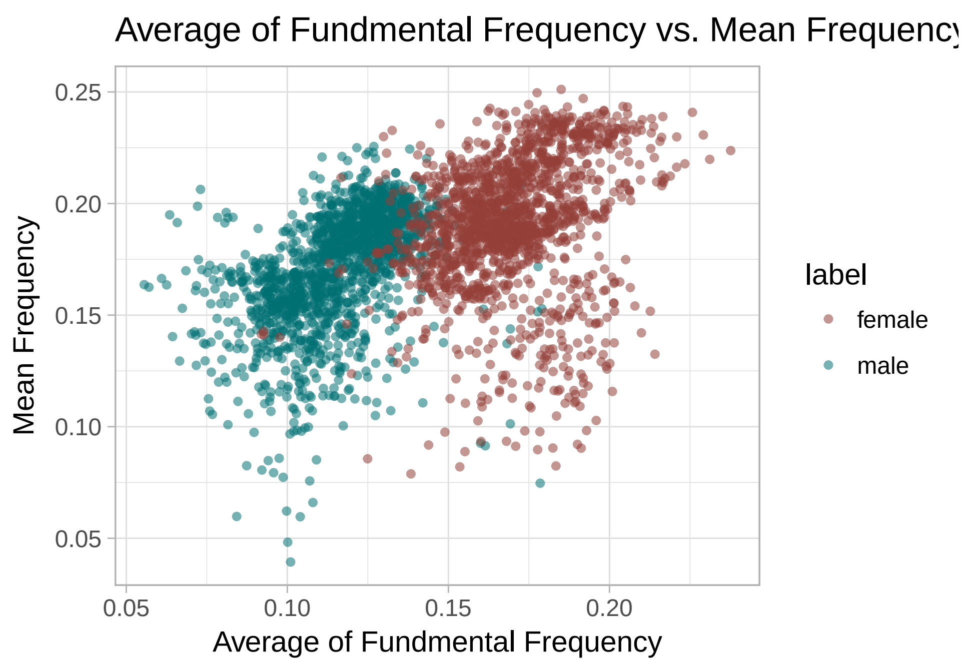

Number of People in Each Gender
1584
1584
1500
1000
500
0
Count
female
male
Gender
label
female
male

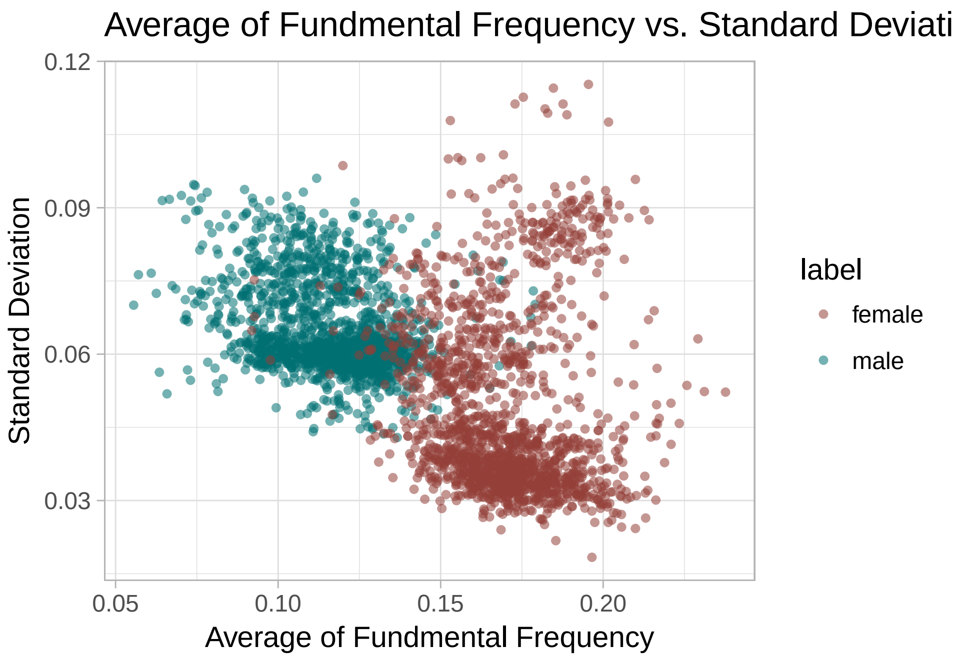

Density of all variables
centroid
dfrange
IQR
kurt
maxdom
maxfun
meandom
meanfreq
meanfun
median
mindom
minfun
mode
modindx
Q25
Q75
sd
sfm
skew
sp.ent
density
feature in each column

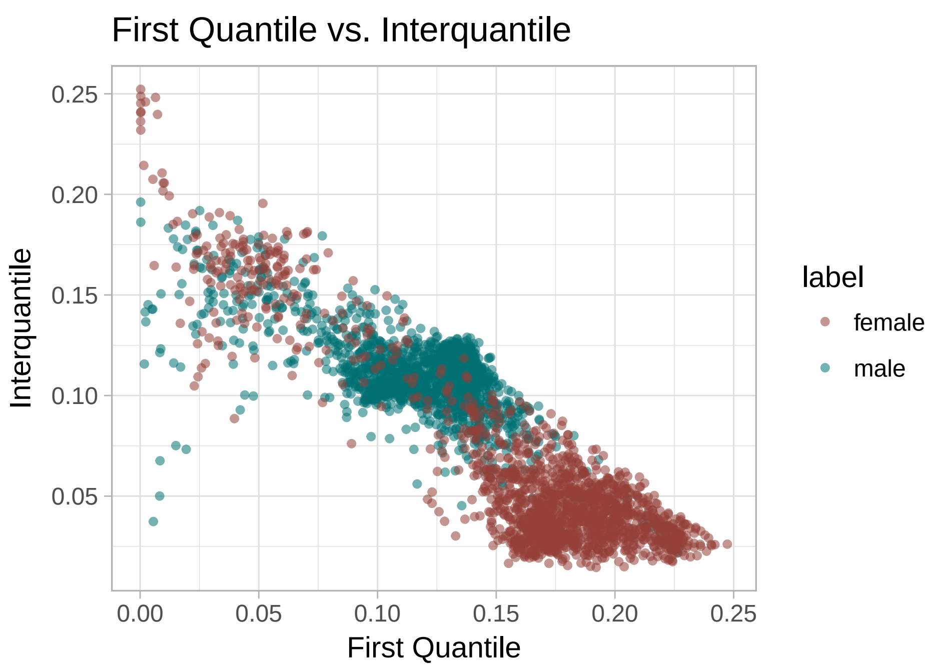

Modulation Index vs. Mean Frequency
Mean Frequency
0.25
0.20
0.15
0.10
0.05
0.00
0.25
0.50
0.75
Modulation Index
label
female
male

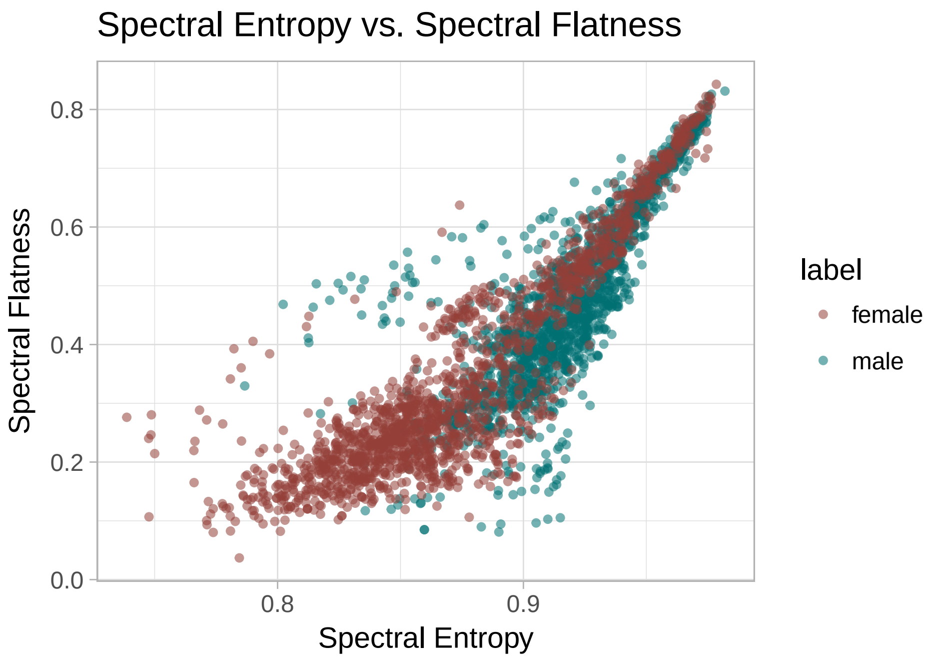

Skew vs. Kurtosis
Kurtosis
1000
500
0
0
10
20
30
Skew
label
female
male


Standard Deviation vs. Mean Frequency
Mean Frequency
Standard Deviation of Frequency
0.25
0.20
0.15
0.10
0.05
0.03
0.06
0.09
0.12
label
female
male


Average of Fundmental Frequency vs. Mean Frequency
Mean Frequency
Average of Fundmental Frequency
0.25
0.20
0.15
0.10
0.05
0.05
0.10
0.15
0.20
label
female
male


Average of Fundmental Frequency vs. Standard Deviati
Standard Deviation
0.12
0.09
0.06
0.03
0.05
0.10
0.15
0.20
Average of Fundmental Frequency
label
female
male


First Quantile vs. Interquantile
Interquantile
0.25
0.20
0.15
0.10
0.05
0.00
0.05
0.10
0.15
0.20
0.25
First Quantile
label
female
male



III. Data Preprocessing

1.Set trainning and testing

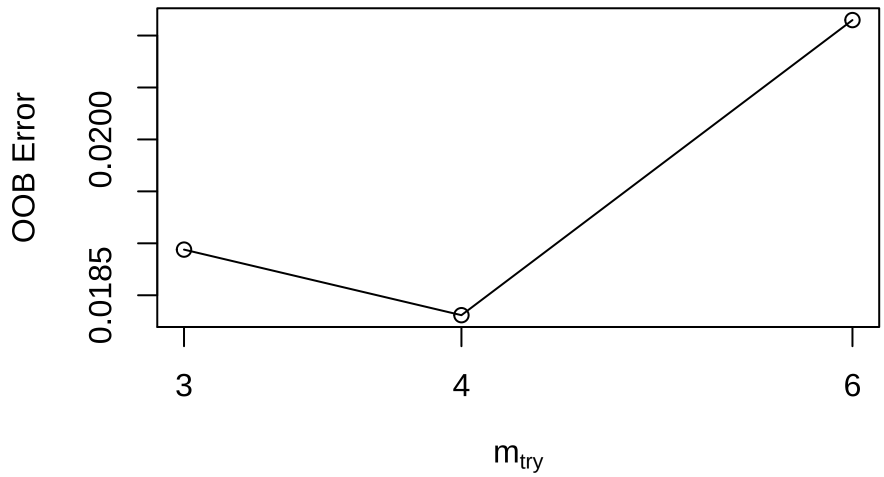
- Randomly select 70% train and 30% test groups
- After feature selections, we will only include selected features in training and testing.

2.Feature selection

Using Random Forest Importance to select features.

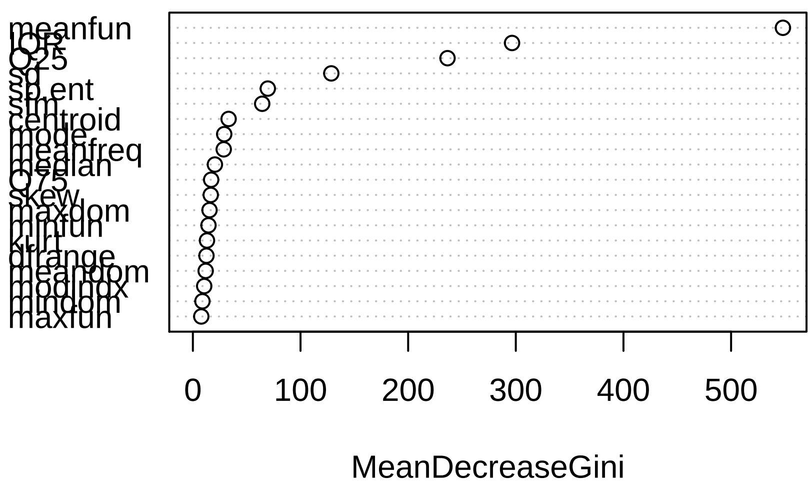
- First, use random forest model as classification
- Second, in order to fit the model, want to find the best mtry (number of variabels selected at each split)
- Third, according to plot of importance in the desending order, we select top 7 important features.

```
mtry = 4  OOB error = 1.83%
Searching left ...
mtry = 3    OOB error = 1.89%
-0.03448276 0.001
Searching right ...
mtry = 6    OOB error = 2.11%
-0.1551724 0.001
```

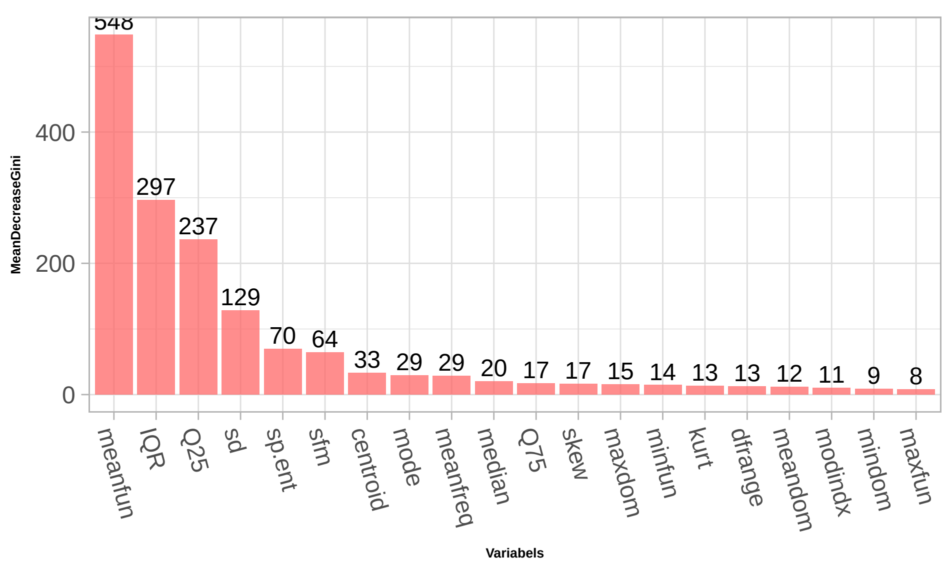



```
[1] "Therefore, based on the plot above, the best number of variables at each split is 4"
```

trainingmodel



Importance of Variables in descending order



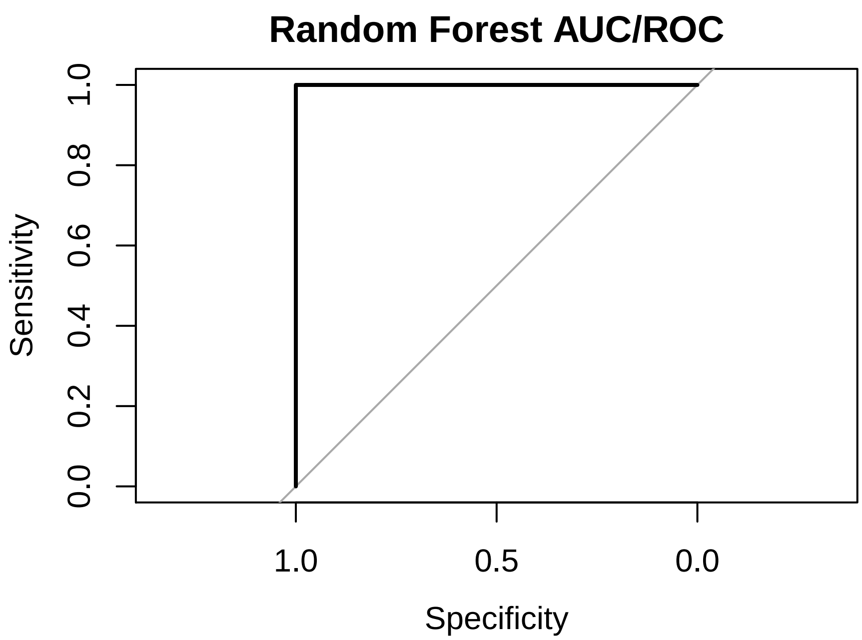
```
[1] "The selected features and the target variable are:"

[1] "sd"       "Q25"      "IQR"      "sp.ent"   "sfm"      "centroid"
[7] "meanfun"  "label"
```

IV. Models Building

1. Random Forest Classification

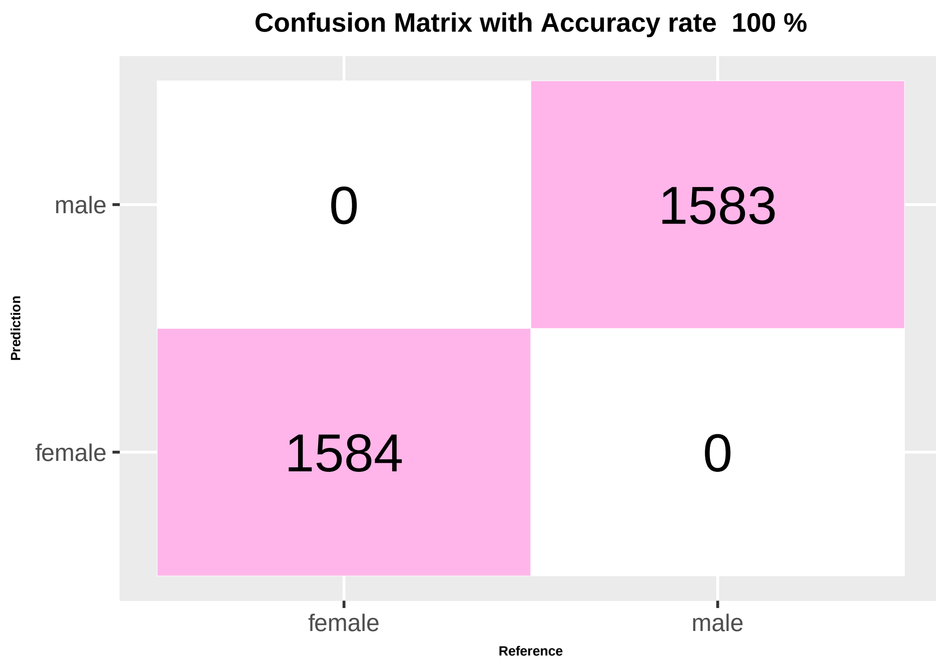
- set parameters for random forest model
- plot the ROC/AUC and confusion matrix

```
           actual
predictions female male
     female   1584    0
     male        0 1583

[1] "In this model, the accuracy rate is:"

[1] 1
```




```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female   1584    0
    male        0 1583

               Accuracy : 1
                 95% CI : (0.9988, 1)
    No Information Rate : 0.5002
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
```

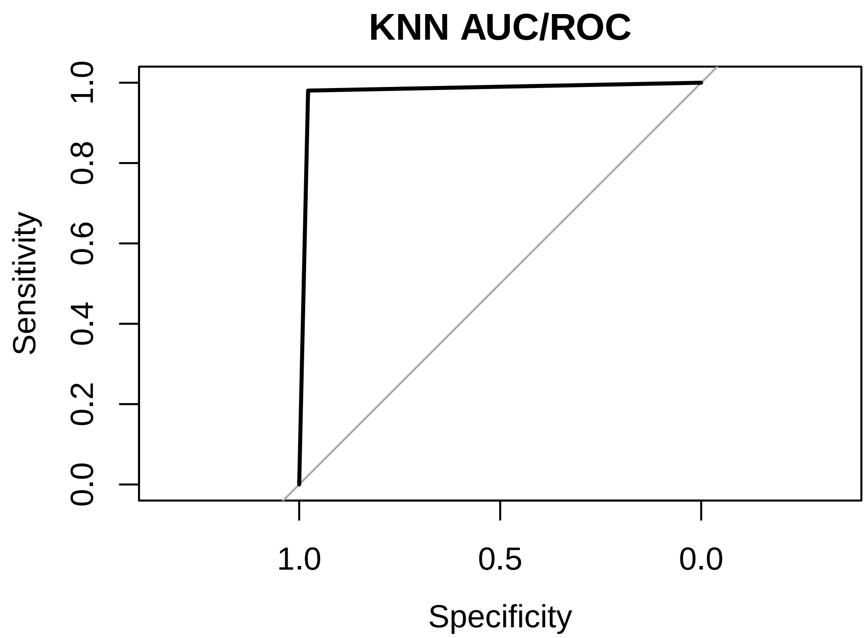
```
 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.5002
         Detection Rate : 0.5002
   Detection Prevalence : 0.5002
      Balanced Accuracy : 1.0000

       'Positive' Class : female
```



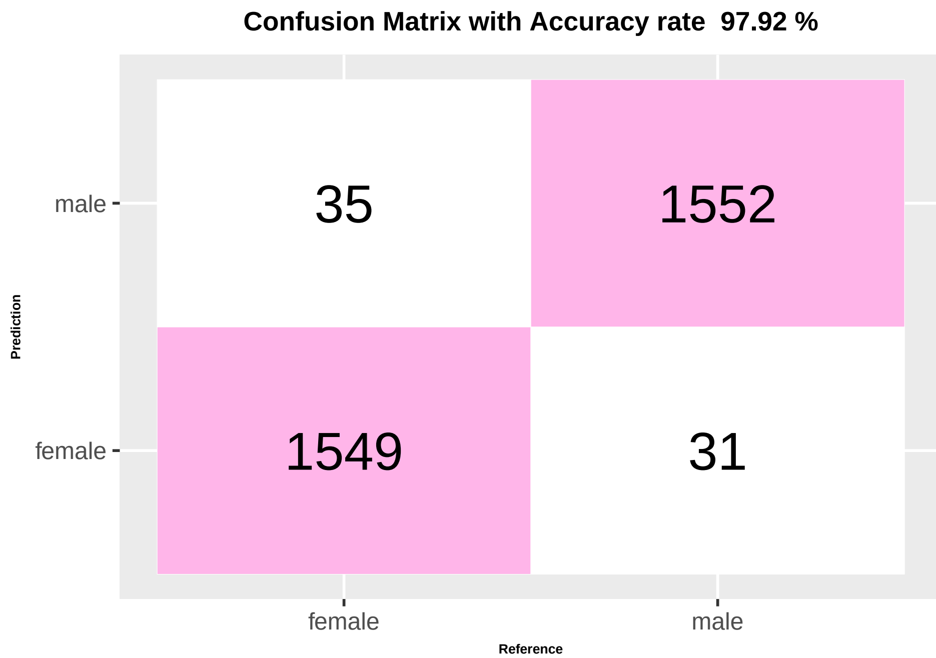
Based on the confusion matrix, blablablabla.....

2. K-Nearest Neighbour classification

- set parameter k = 7
- plot confusion matrix




```
[1] "The accuracy rate in KNN is:"

[1] 0.9791601

Confusion Matrix and Statistics

          Reference
Prediction female male
    female   1549   31
    male       35 1552

               Accuracy : 0.9792
                 95% CI : (0.9736, 0.9838)
    No Information Rate : 0.5002
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9583
 Mcnemar's Test P-Value : 0.7119

            Sensitivity : 0.9779
            Specificity : 0.9804
         Pos Pred Value : 0.9804
         Neg Pred Value : 0.9779
             Prevalence : 0.5002
         Detection Rate : 0.4891
   Detection Prevalence : 0.4989
      Balanced Accuracy : 0.9792

       'Positive' Class : female
```

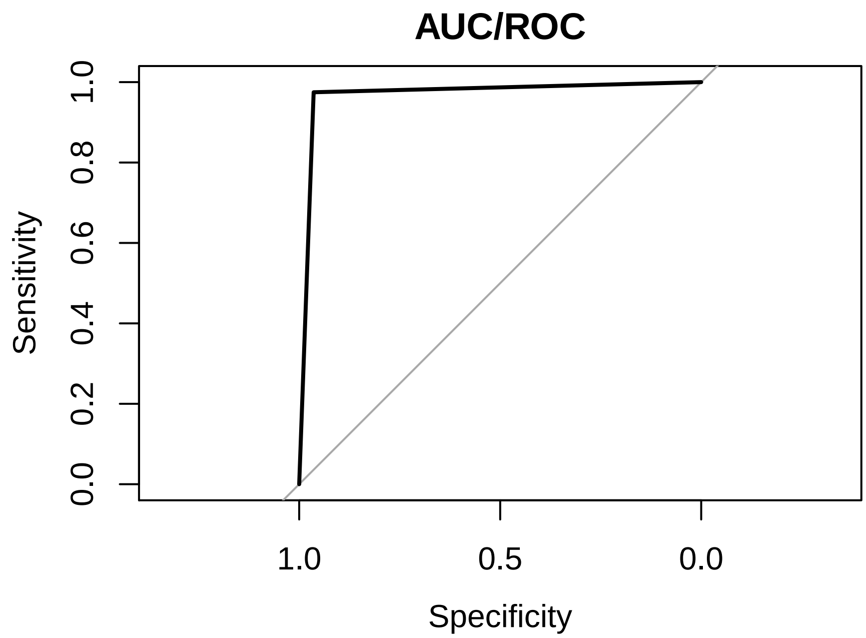


3. Logistic Regression

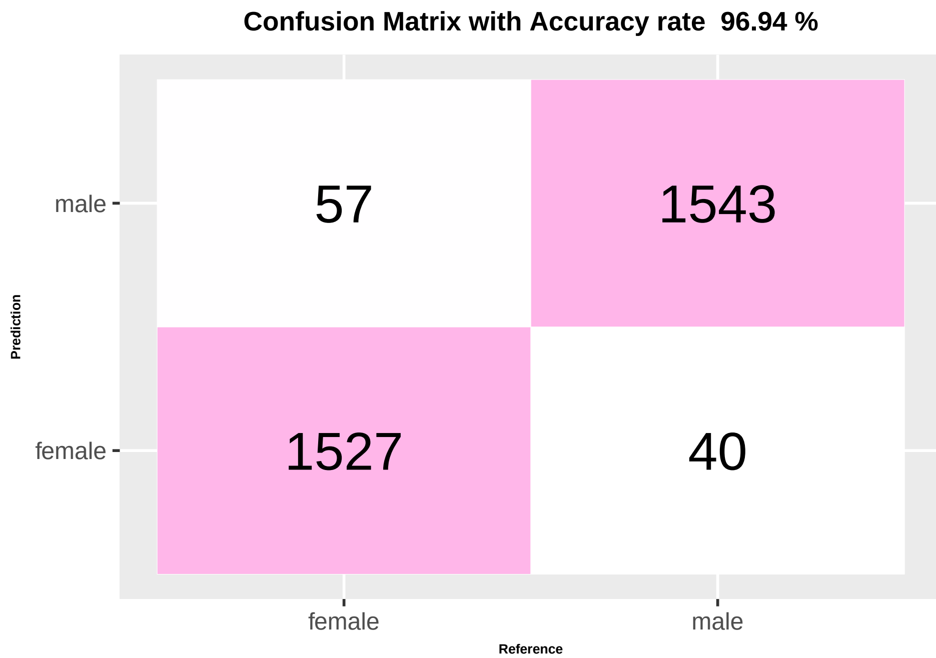
```
Call:
glm(formula = label ~ ., family = binomial(link = "logit"), data = train,
    control = list(maxit = 50))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0803  -0.0396   0.0002   0.1112   4.2916

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -18.838      6.856  -2.748    0.006 **
sd           -30.919     26.101  -1.185    0.236
Q25            1.776     16.654   0.107    0.915
IQR           66.337     13.164   5.039 4.67e-07 ***
sp.ent        45.842      7.838   5.848 4.96e-09 ***
sfm          -11.842      2.406  -4.922 8.58e-07 ***
centroid       3.525     19.640   0.179    0.858
meanfun     -161.144      8.231 -19.578  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4391.78  on 3167  degrees of freedom
Residual deviance:  610.18  on 3160  degrees of freedom
AIC: 626.18

Number of Fisher Scoring iterations: 8
```

```
           actual
predictions female male
          0   1527   40
          1     57 1543
```

```
[1] 0.9693716
```




```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female   1527   40
    male       57 1543

               Accuracy : 0.9694
                 95% CI : (0.9628, 0.9751)
    No Information Rate : 0.5002
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9387
 Mcnemar's Test P-Value : 0.1043

            Sensitivity : 0.9640
            Specificity : 0.9747
         Pos Pred Value : 0.9745
         Neg Pred Value : 0.9644
             Prevalence : 0.5002
         Detection Rate : 0.4822
   Detection Prevalence : 0.4948
      Balanced Accuracy : 0.9694

       'Positive' Class : female
```



V. Conclusion

-The accuracy of all three models is over 96%

-Gender can be recognized by voice. We have demo to show the gender recognition process during our presentation in class.

-After finish the major parts of the project, we are still curious about whether people???s disguised voice can be recognized or not. If we add some feigned voices into the dataset, we might get some different results