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 voice of a person contains many different properties, and these properties often reflect the certain information about the person, such as gender. The objective of this research is to build a gender recognition system based on different classification algorithms. The dataset includes the measurement of each voice sample's auditory features which are based upon acoustic properties of the voice. The voice samples are pre-processed by acoustic analysis in R using the seewave and tuneR packages.

II. Data information & Exploratory Data Analysis

1. Data Information

The dataset consists of acoustic properties, such as frequencies, of audio samples generated by human voices. The focus of the project to use these properties to accurately predict gender based on voice.

The source of the data is from this website: **Primary Objects** by *KORY BECKER*. In this dataset, it contains 3169 voice samples and 21 columns (our target variable is `label`).

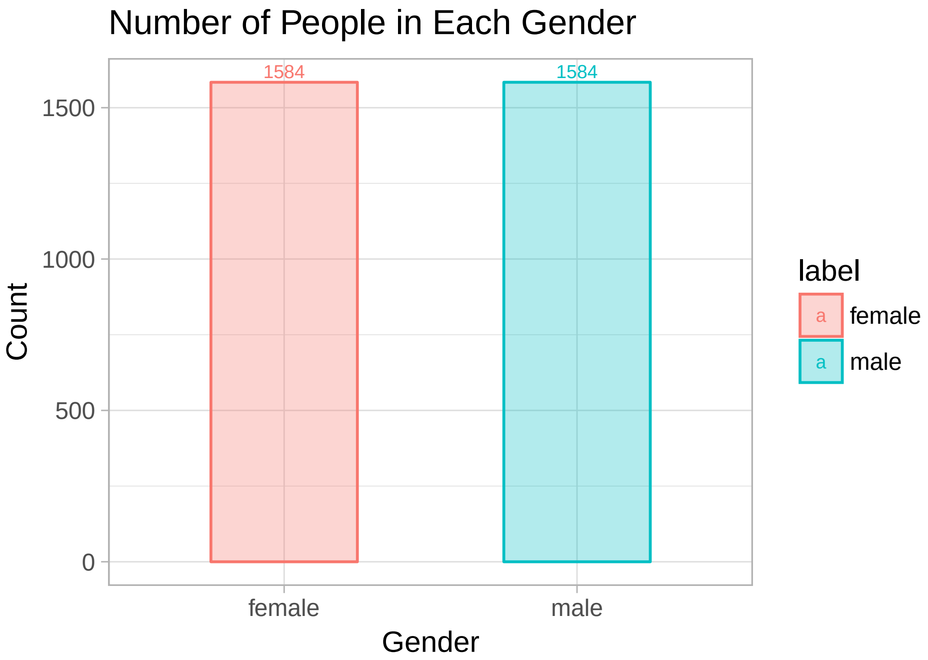
- duration: length of signal
- meanfreq: mean frequency (in kHz)
- sd: standard deviation of frequency
- median: median frequency (in kHz)
- Q25: first quantile (in kHz)
- Q75: third quantile (in kHz)
- IQR: interquantile range (in kHz)
- skew: skewness (see note in specprop description)
- kurt: kurtosis (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)

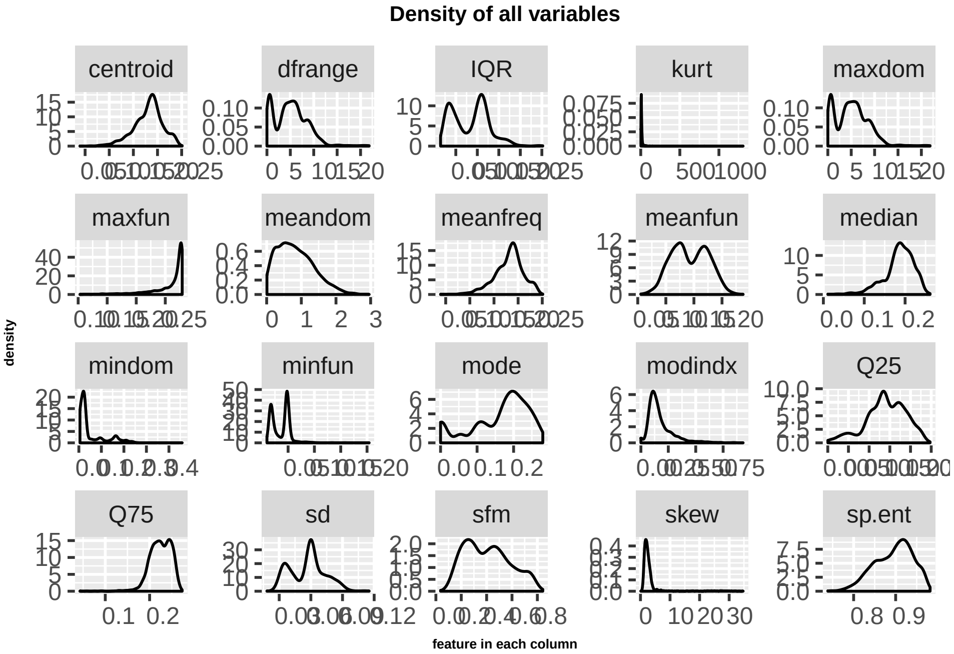
Three classification models were considered for gender recognition:

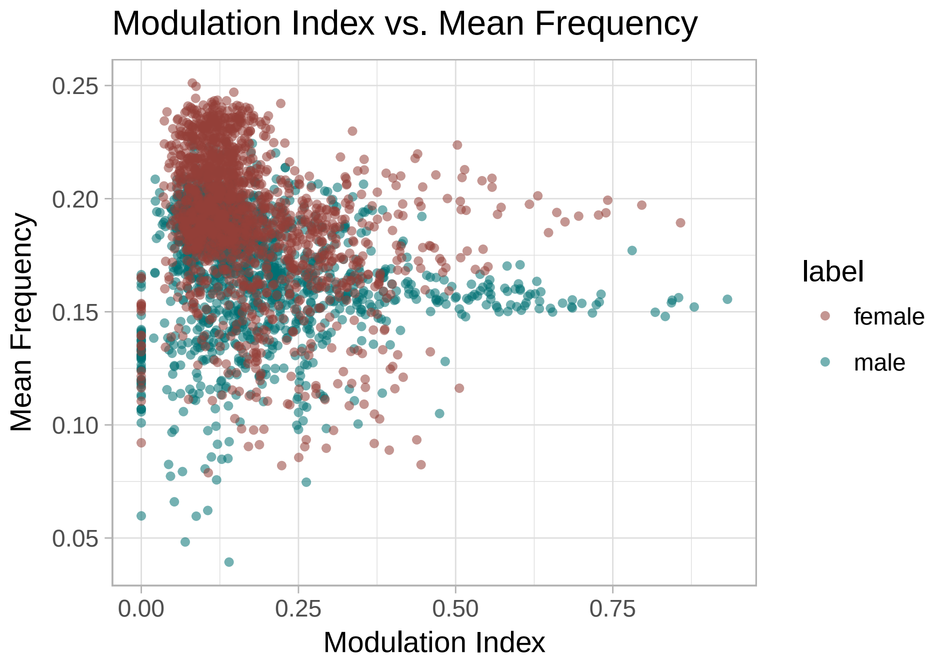
1. Random Forest
2. K-Nearest Neighbour
3. Logistic Regression

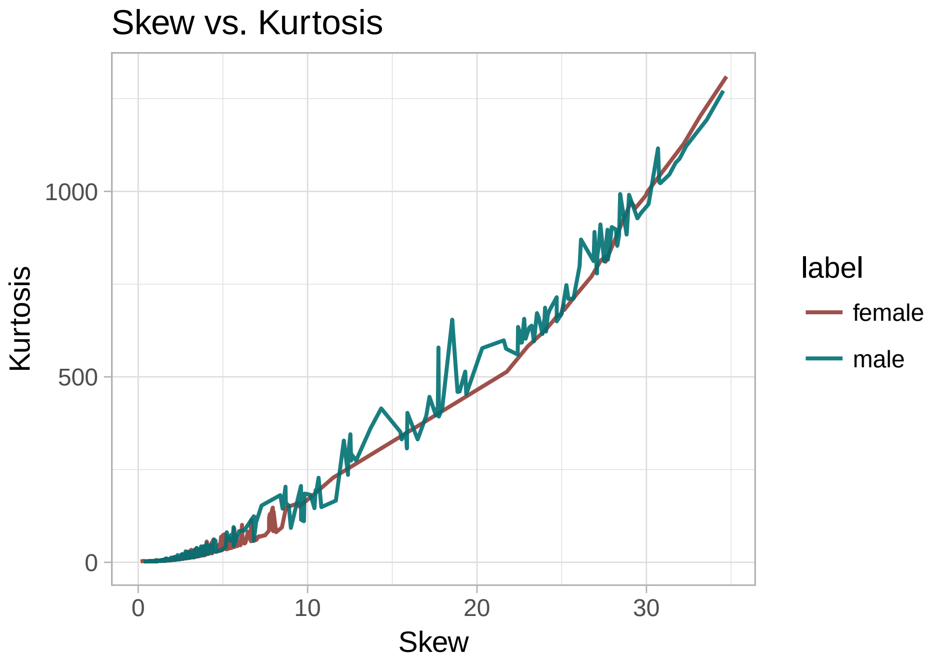
Finally, the confusion matricies and highest accuracies will be used to determine the best classification model.

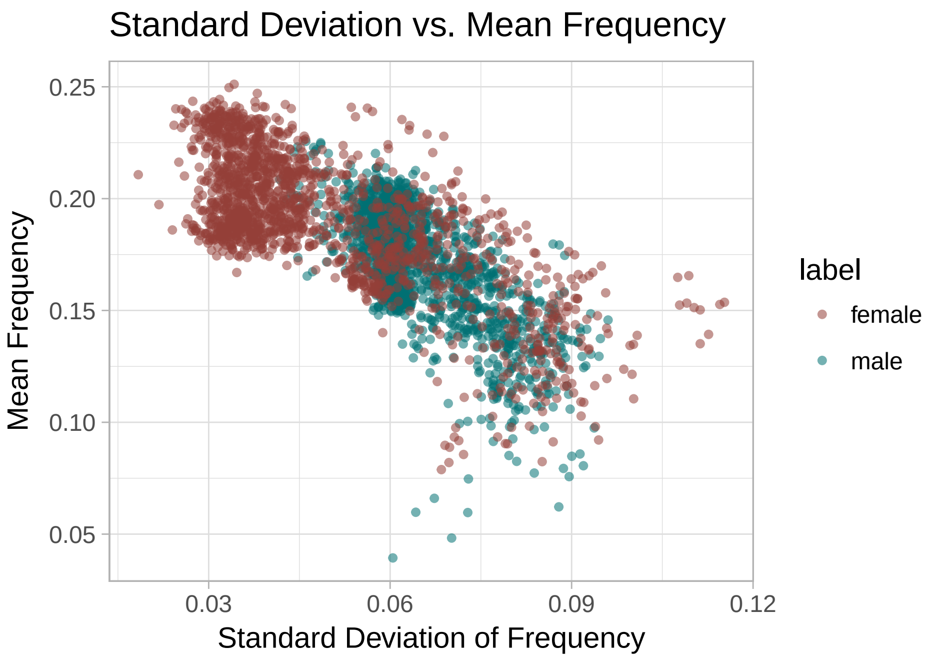
2. Exploratory Data Analysis

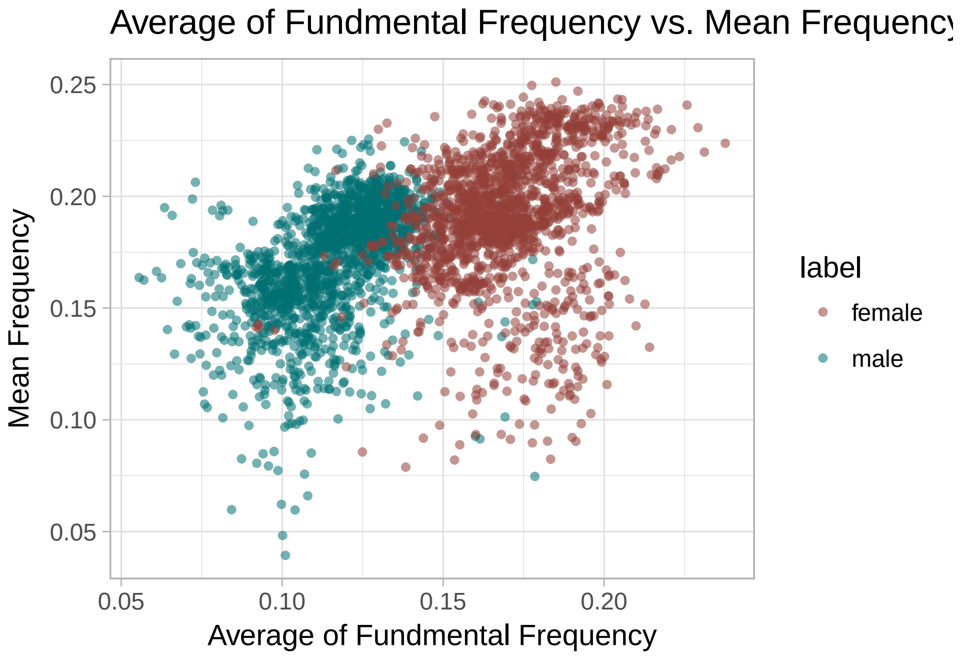


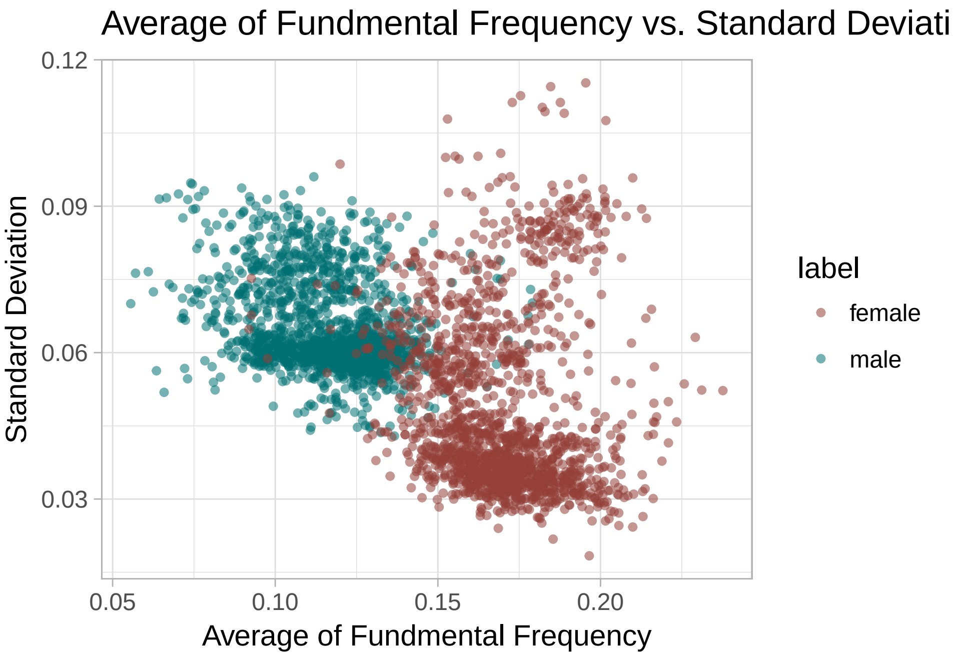



Modulation Index vs. Mean Frequency
Mean Frequency
0.25
0.20
0.15
0.10
0.05
0.00
0.25
0.50
0.75
Modulation Index
label
female
male

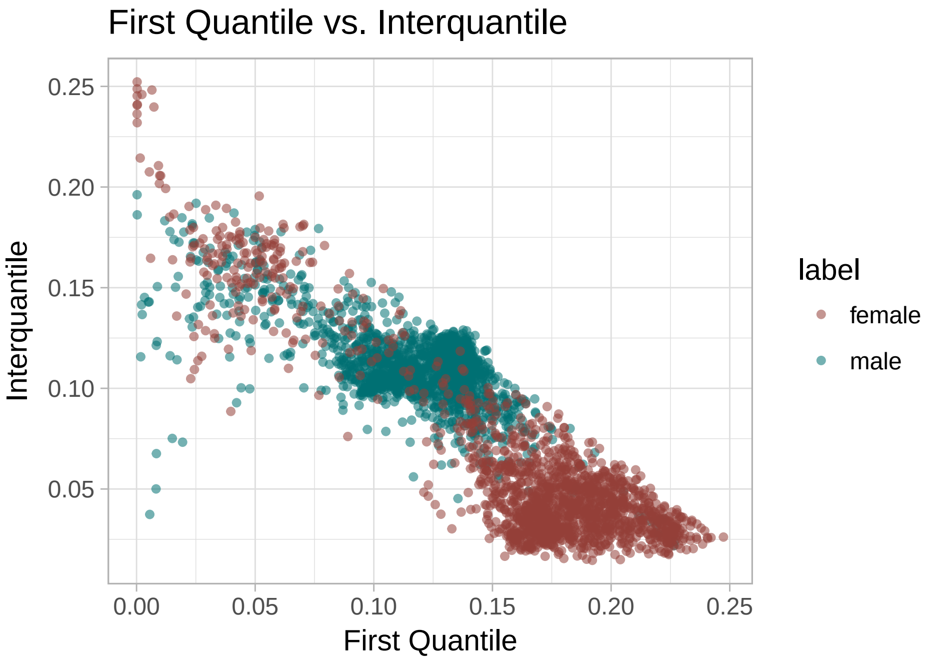

Skew vs. Kurtosis
Kurtosis
1000
500
0
0
10
20
30
Skew
label
female
male


Standard Deviation vs. Mean Frequency

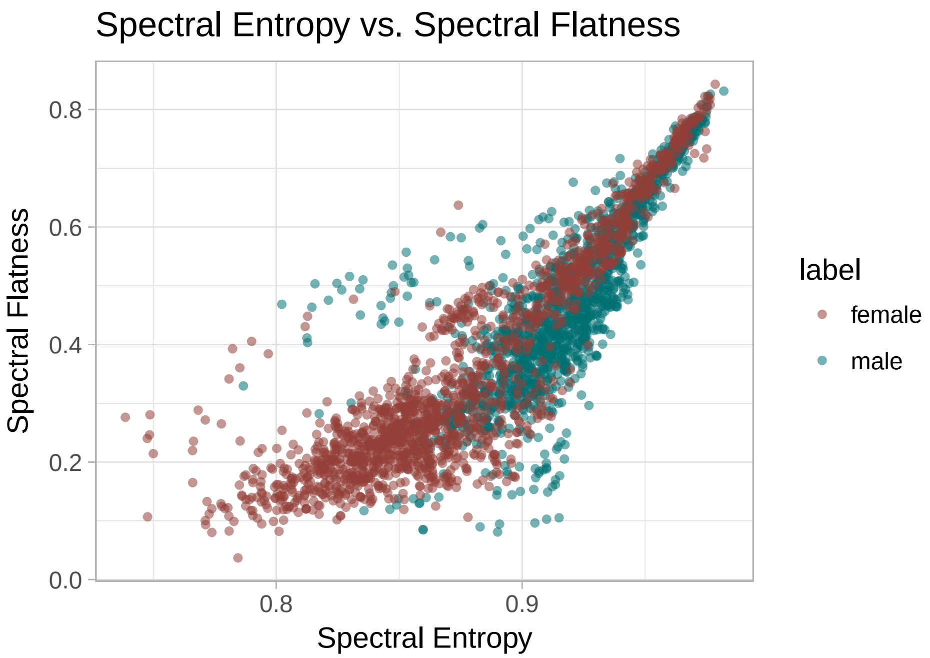

Average of Fundmental Frequency vs. Mean Frequency



Average of Fundmental Frequency vs. Standard Deviati



First Quantile vs. Interquantile



III. Data Preprocessing

1.Set training 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.

Forest error rate depends on two things:

1. The correlation between any two trees in the forest. Increasing the correlation increases the forest error rate.
2. The strength of each individual tree in the forest.

A tree with a low error rate is a strong classifier. Therefore, we want to increase the strength of the individual trees and decrease decreases the forest error rate.

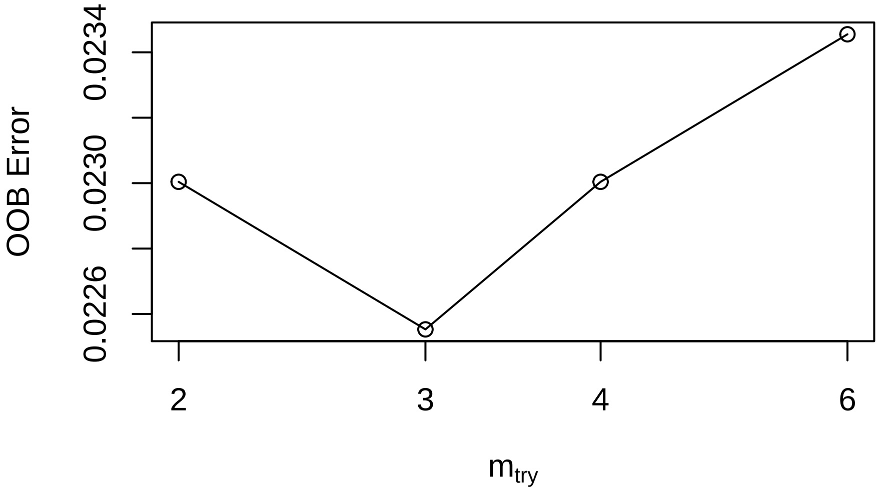
However, reducing m reduces both the correlation and the strength. Increasing it increases both. Somewhere in between is an "optimal" range of mtry - usually quite wide. Using the oob error rate (see the plot below) can give a value of m in the range can quickly be found.

Therefore, for feature selection, we need to do two steps:

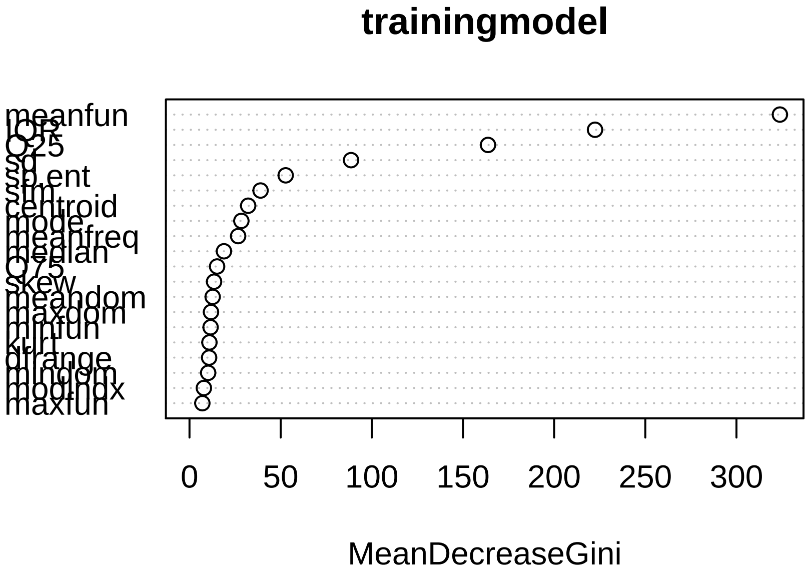
1. Find the best mtry (number of variables selected at each split)
2. According to plot of importance in the desending order, we select top 7 important features.

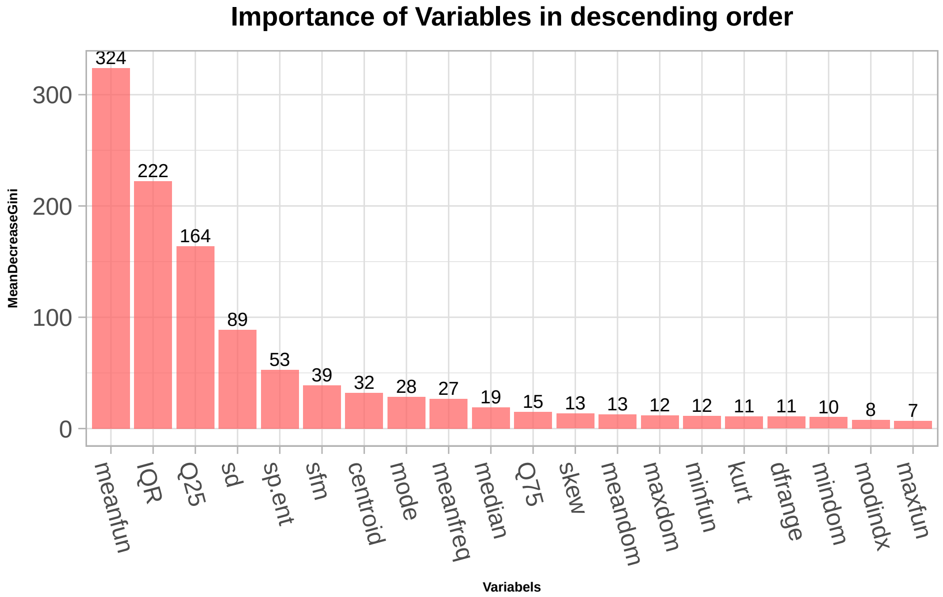
```
mtry = 4  OOB error = 2.3%
Searching left ...
mtry = 3    OOB error = 2.26%
0.01960784 0.001
```

```
mtry = 2    OOB error = 2.3%
-0.02 0.001
Searching right ...
mtry = 6    OOB error = 2.35%
-0.04 0.001
```



```
[1] "Therefore, based on the plot above, the best number of variables at each split is 3"
```





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

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

IV. Models Building

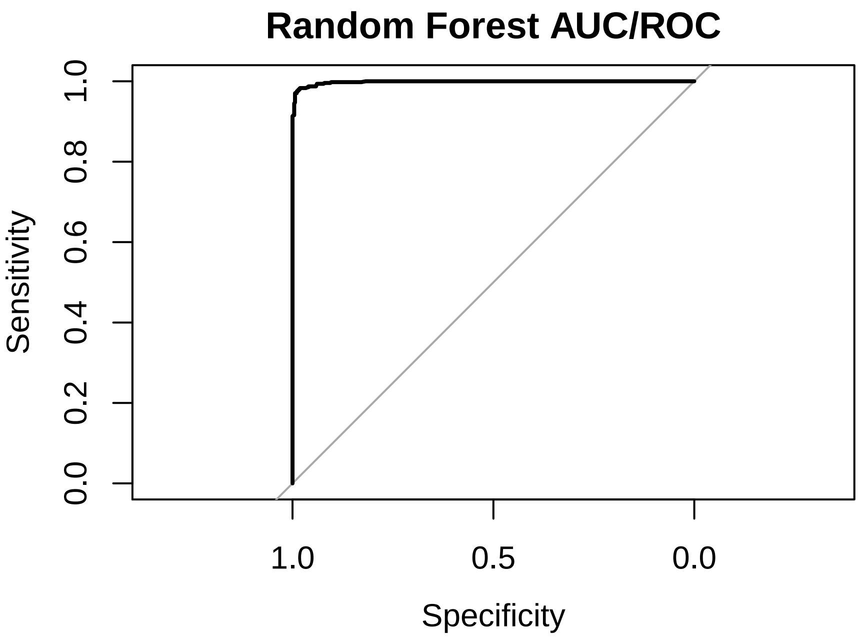
1. Random Forest Classification

- Set parameters for random forest model
- Plot the ROC/AUC and confusion matrix

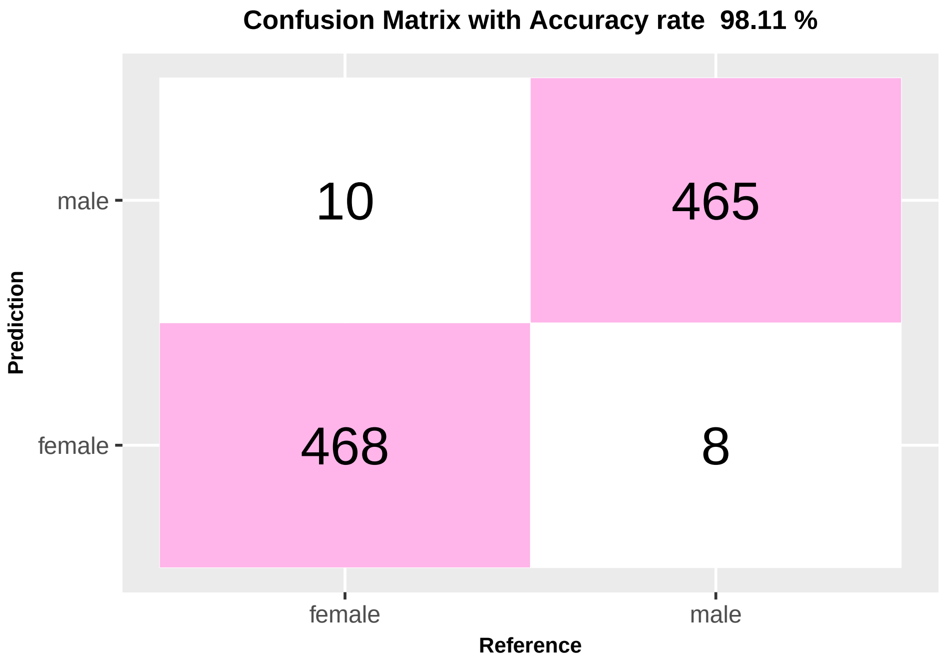
```
           actual
predictions female male
     female    468    8
     male       10  465
```

```
[1] "In the Random Forest model, the accuracy rate is:"
```

```
[1] "98.11 %"
```



```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    468    8
    male       10  465

               Accuracy : 0.9811
                 95% CI : (0.9703, 0.9887)
    No Information Rate : 0.5026
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9621
 Mcnemar's Test P-Value : 0.8137

            Sensitivity : 0.9791
            Specificity : 0.9831
         Pos Pred Value : 0.9832
         Neg Pred Value : 0.9789
             Prevalence : 0.5026
         Detection Rate : 0.4921
   Detection Prevalence : 0.5005
      Balanced Accuracy : 0.9811

       'Positive' Class : female
```



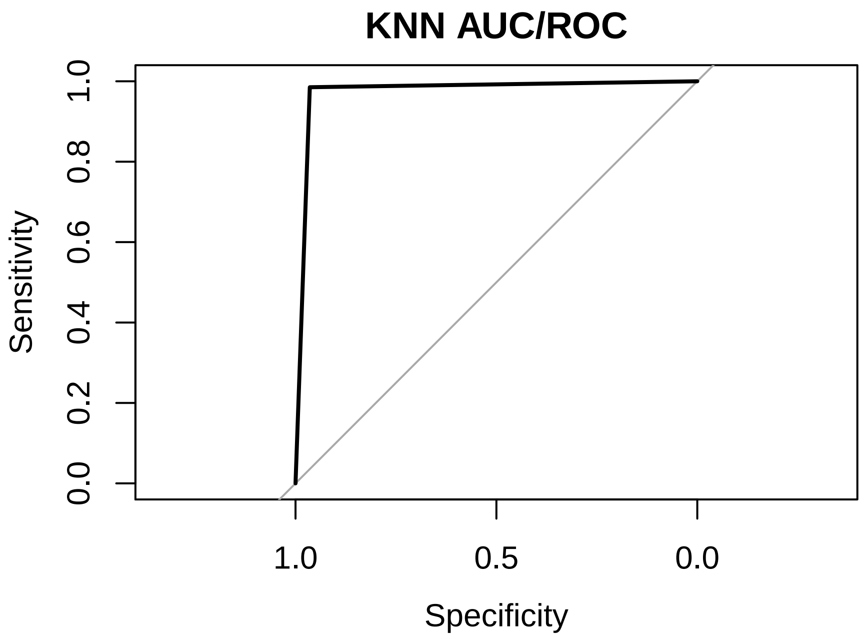
Based on the confusion matrix, there are 468 samples are predicted correctly as females; however, there are 10 samples predicted as males but are actually females. There are 465 samples that are predicted correctly as males; however, there are 8 samples that are predicted as female but are in fact male.

Based on the elbow method, we find that the first segment of line in AUC is almost parallel to the y-axis and the angel of the left corner is nearly 90 degree, which means that the area under curve is close 100, and also means that this is a good model.

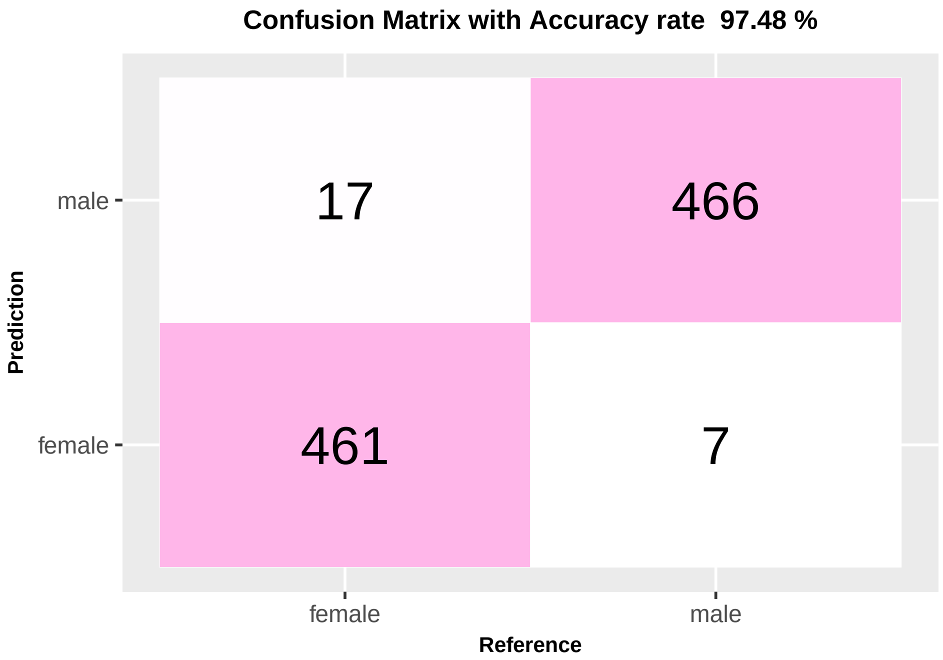
The accuracy rate is 98.11% and it is in the 95% confidence interval with the p-value far smaller than 0.05. Therefore, it can be concluded that this accuracy is statistically significant.

2. K-Nearest Neighbour classification

- Set parameter k = 7. When k=7, compared to other k values, the accuracy reach the highest.
- Plot ROC/AUC and confusion matrix



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

[1] "97.48 %"

Confusion Matrix and Statistics

          Reference
Prediction female male
    female    461    7
    male       17  466

               Accuracy : 0.9748
                 95% CI : (0.9627, 0.9838)
    No Information Rate : 0.5026
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.9495
 Mcnemar's Test P-Value : 0.06619

            Sensitivity : 0.9644
            Specificity : 0.9852
         Pos Pred Value : 0.9850
         Neg Pred Value : 0.9648
             Prevalence : 0.5026
         Detection Rate : 0.4848
   Detection Prevalence : 0.4921
      Balanced Accuracy : 0.9748

       'Positive' Class : female
```



Based on the confusion matrix, there are 461 samples are predicted correctly as females; however, there are 17 samples predicted as males but are actually females. There are 466 samples that are predicted correctly as males; however, there are 7 samples that are predicted as female but are in fact male.

The accuracy is 97.48% and is in the 95% confidence interval with the p-value far smaller than 0.05. Therefore, we can say that this accuracy is statistically significant.

3. Logistic Regression

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9548  -0.0493   0.0013   0.1191   4.2098

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -13.953      8.045  -1.734  0.08286 .
sd           -36.689     30.200  -1.215  0.22441
Q25           -3.632     19.253  -0.189  0.85035
IQR           60.964     15.289   3.987 6.68e-05 ***
sp.ent        39.625      9.044   4.381 1.18e-05 ***
sfm          -10.443      2.755  -3.791  0.00015 ***
centroid       9.733     22.989   0.423  0.67202
meanfun     -156.511      9.391 -16.667  < 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: 3073.40  on 2216  degrees of freedom
Residual deviance:  448.89  on 2209  degrees of freedom
AIC: 464.89

Number of Fisher Scoring iterations: 8
```

Based on the summary table above, the P-values of `sd`, `Q25` and `centroid` are larger than 0.05, which means those variables are not significant. Therefore, those variables need to be remove in the next model.

```
Call:
glm(formula = label ~ IQR + sp.ent + sfm + meanfun, family = binomial(link = "logit"),
    data = train, control = list(maxit = 50))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9738  -0.0448   0.0014   0.1215   4.1803

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -20.516      6.418  -3.197  0.00139 **
IQR           55.111      4.826  11.419  < 2e-16 ***
sp.ent        47.884      7.512   6.375 1.83e-10 ***
sfm          -14.038      1.743  -8.054 8.00e-16 ***
meanfun     -155.123      9.271 -16.732  < 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: 3073.40  on 2216  degrees of freedom
Residual deviance:  451.62  on 2212  degrees of freedom
AIC: 461.62

Number of Fisher Scoring iterations: 8
```

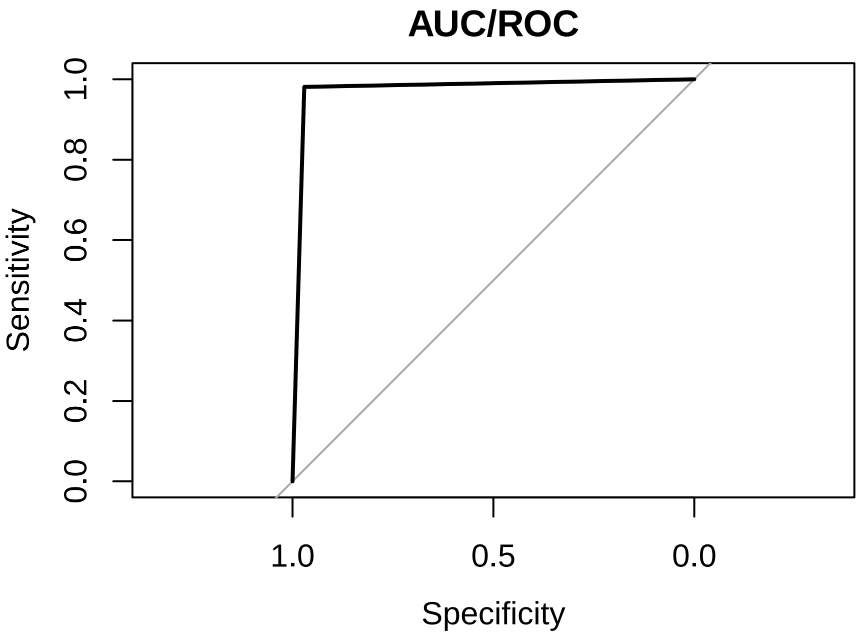
From the second summary:

1. all the variables in the table above are very sinificant since the p-values are far smaller than 0.05.
2. For every one unit increase in `IQR`, log odds of gender as female increases by 55.11; For every one unint increase in `sp.ent`,the log odds of fender as female increase by 47.884. For every one unit increase in `sfm`, the log odds of gender as female decrease -14.038 (more likely to be male). For every one unit increase in `meanfun`, the log odds of gender as female decrease 155.123 (more likely to be male).
3. The Residual deviance increase from 448.89 to 451.62, which means the model is fitting good.
4. The Deviance Residuals is symmetrically distributed at center, a little bit skewed to the right.
5. the AIC dropped from 464.89 to 461.62, which means the this model is fitting good as well.

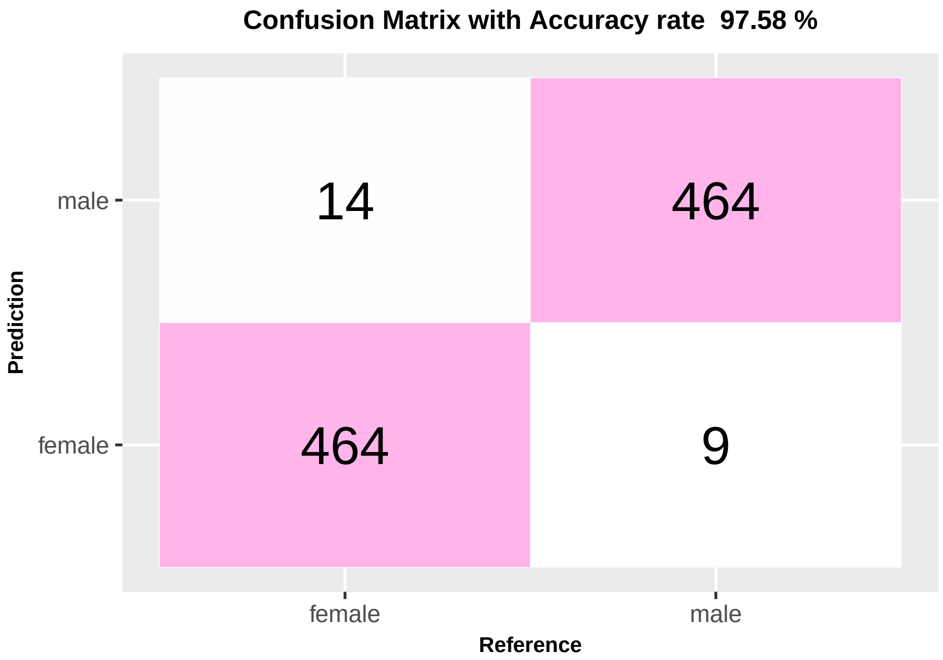
```
           actual
predictions female male
          0    464    9
          1     14  464
```

```
[1] "The accuracy rate in Logistic Regression is:"
```

```
[1] "97.58 %"
```



```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    464    9
    male       14  464

               Accuracy : 0.9758
                 95% CI : (0.9639, 0.9846)
    No Information Rate : 0.5026
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9516
 Mcnemar's Test P-Value : 0.4042

            Sensitivity : 0.9707
            Specificity : 0.9810
         Pos Pred Value : 0.9810
         Neg Pred Value : 0.9707
             Prevalence : 0.5026
         Detection Rate : 0.4879
   Detection Prevalence : 0.4974
      Balanced Accuracy : 0.9758

       'Positive' Class : female
```



Based on the confusion matrix, there are 464 samples that are predicted correctly as females; however, there are 14 samples that are predicted as males but are actually females. There are 464 samples that are predicted correctly as males; however, there are 9 samples are predicted as females but are in fact males.

The accuracy is 97.58% and is in the 95% confidence interval with the p-value far smaller than 0.05. Therefore, this accuracy is statistically significant.

V. Conclusion

- The accuracy rates of all three models are over 97%, all the models performed well on prediction. Among these three, the random forest classification is the best, since its accuracy reaches 98.11%. Therefore, random forest best model for gender recognition.
- The reason why all models have such high accuracy is that the input dataset is perfectly balanced and has no missing values. If our dataset is imbalanced or has many missing values, we need to rebalance the data and also do the imputation, which will decrease the performance of madels.
- Gender can be recognized by voice. We did three recognitions in demo. For the first trial, the person spoke too short, and there were some speaker' noise during recording, the result was false. The rest of two were correct.
- After finishing the major parts of the project, we are still curious that what we can do next to make our machine smarter. What if people disguise their voice? What if we add some feigned voices into the dataset? Can this machine still do a good job?
- For feature selection, there might be some better algorithms to perform dimention reduction such as principal component analysis (PCA) or independent component analysis (ICA).
- In validation process, we only randomly select test for one time. It's more appropriate if we did 10-fold cross validation. Then, the accuracy rate will be more 'accurate'.