

# 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 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

Our dataset is about the voice of gender, therefore, we conduct a question 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 find the data on the 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)

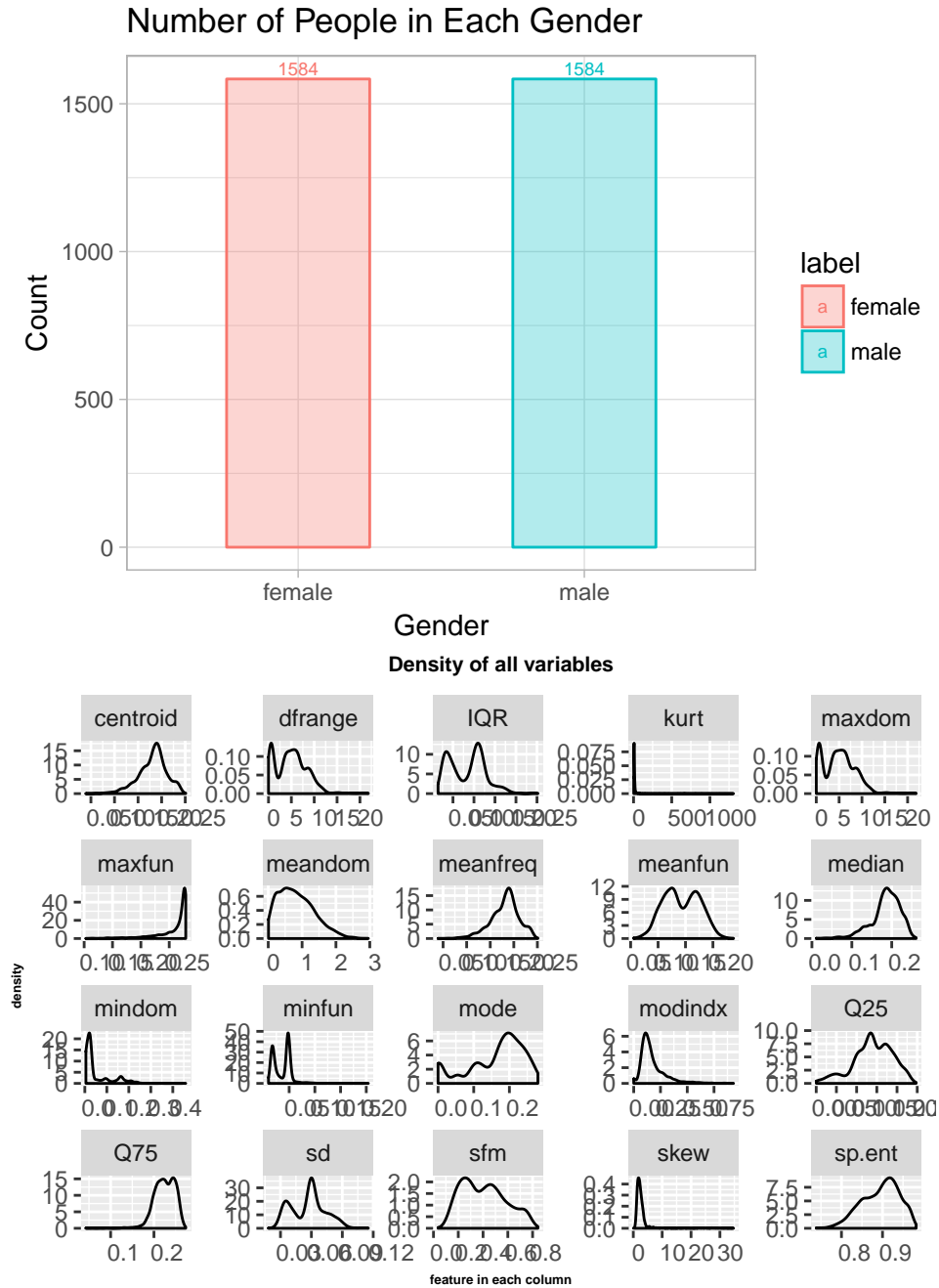
We want to use three classification models to do the gender recognition:

1. Random Forest
2. K-Nearest Neighbour

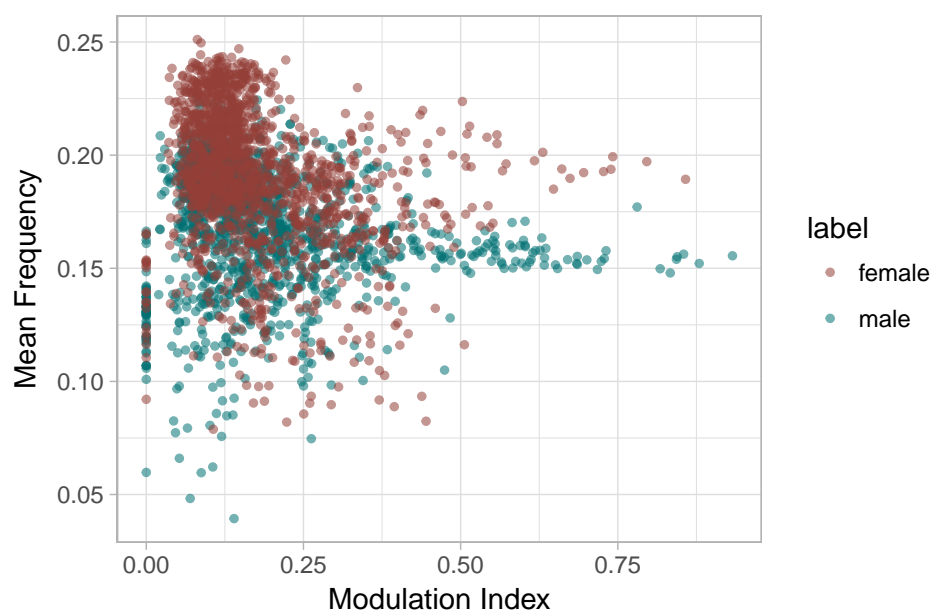
### 3. Logistic Regression

Finally, We use confusion matrix and highest accuracy rate to determine which classification model is the best.

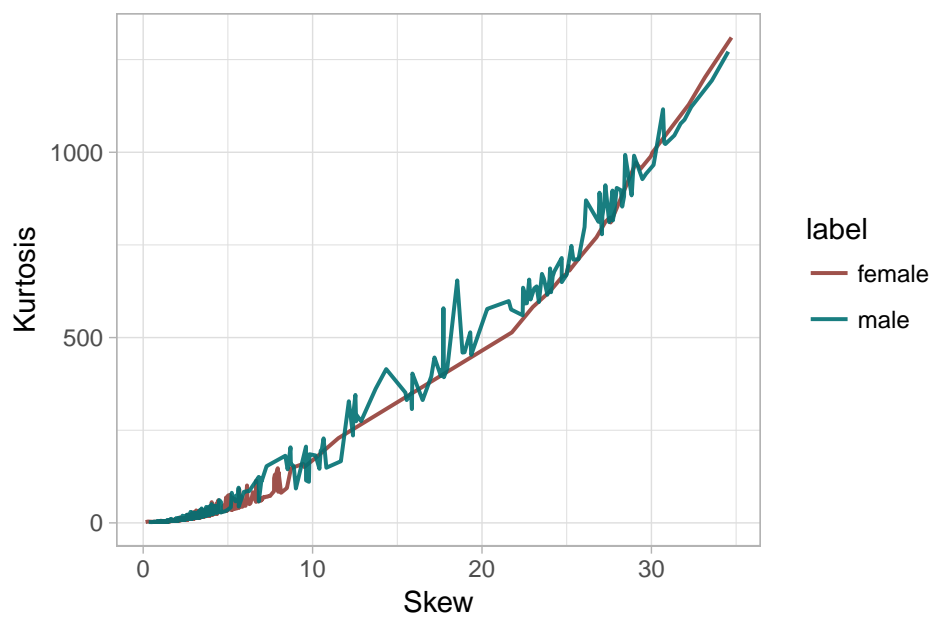
## 2. Exploratory Data Analysis



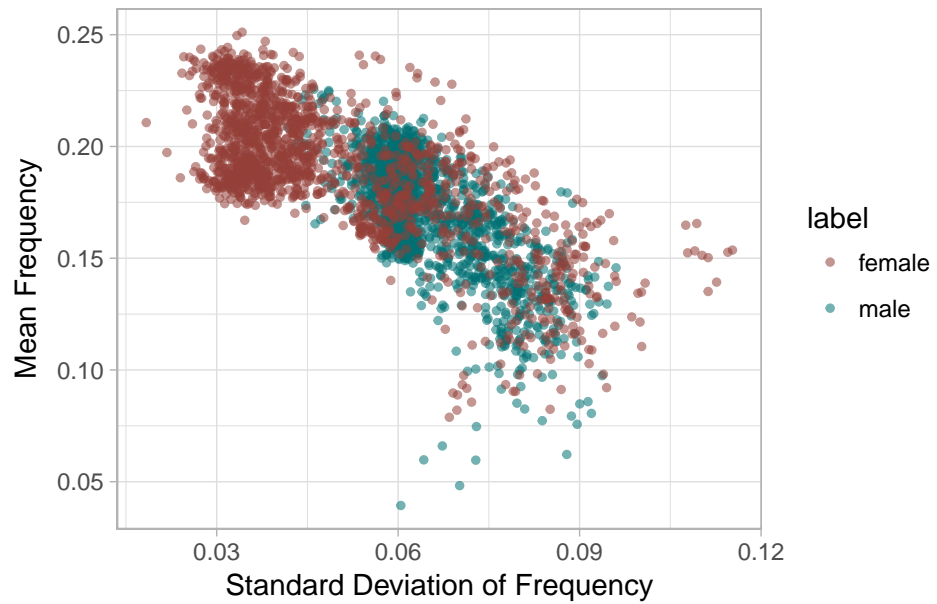
Modulation Index vs. Mean Frequency



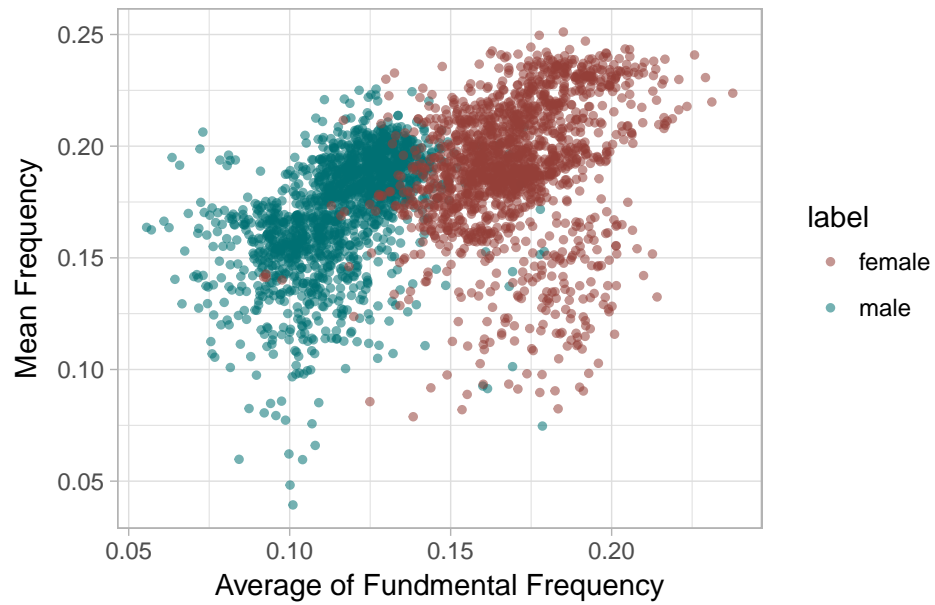
Skew vs. Kurtosis



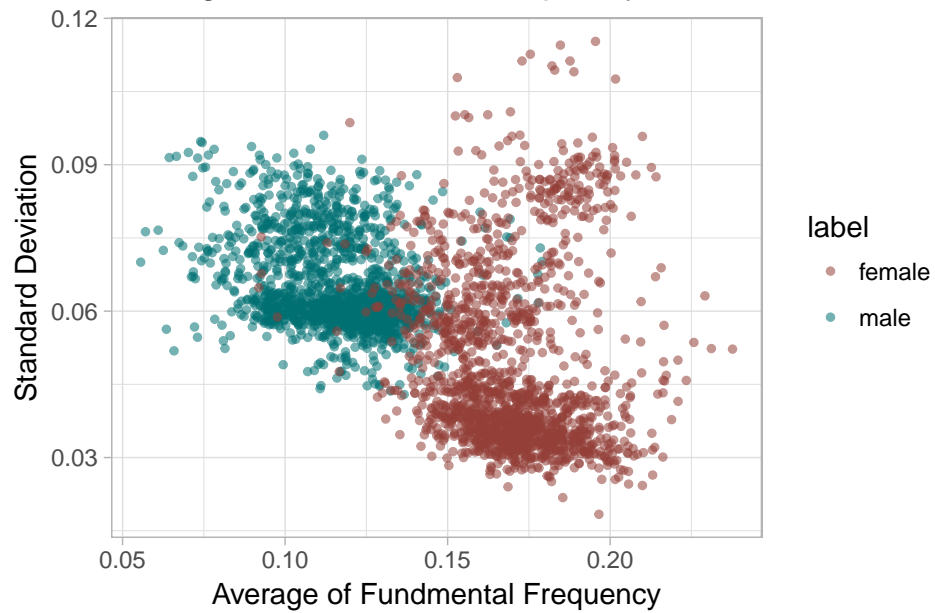
Standard Deviation vs. Mean Frequency



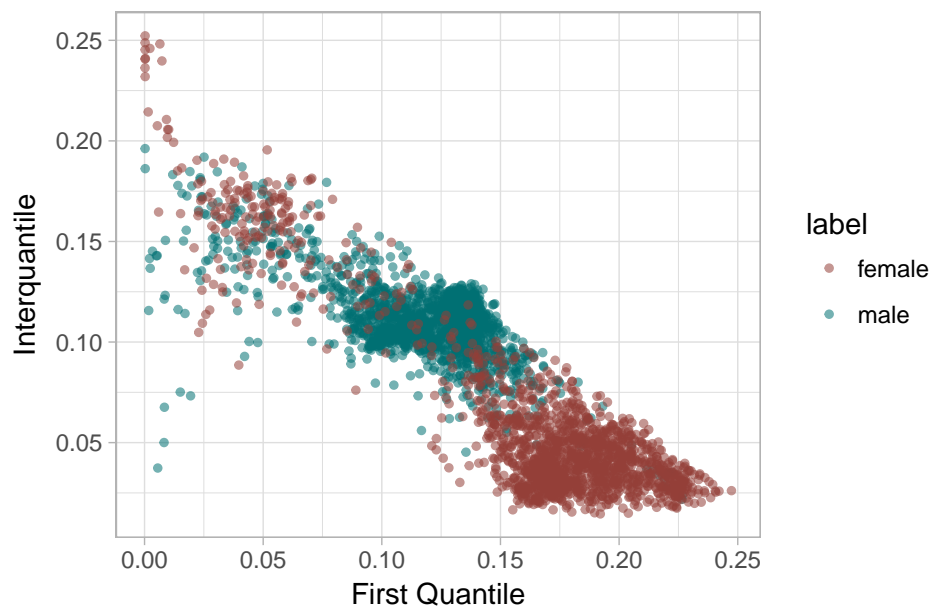
Average of Fundamental Frequency vs. Mean Frequency

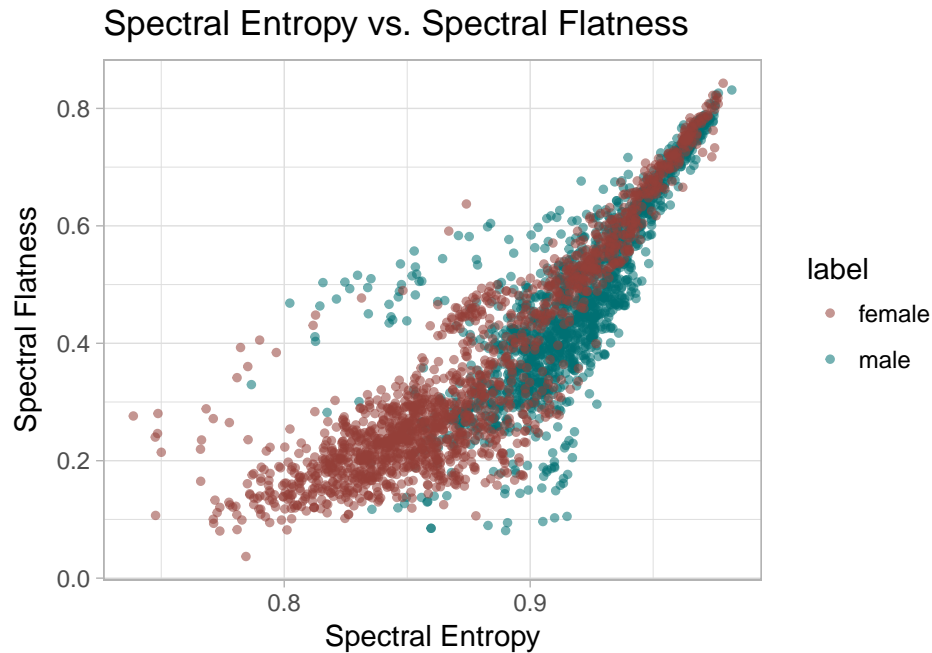


Average of Fundamental Frequency vs. Standard Deviation



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 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  $m_{try}$  - 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:

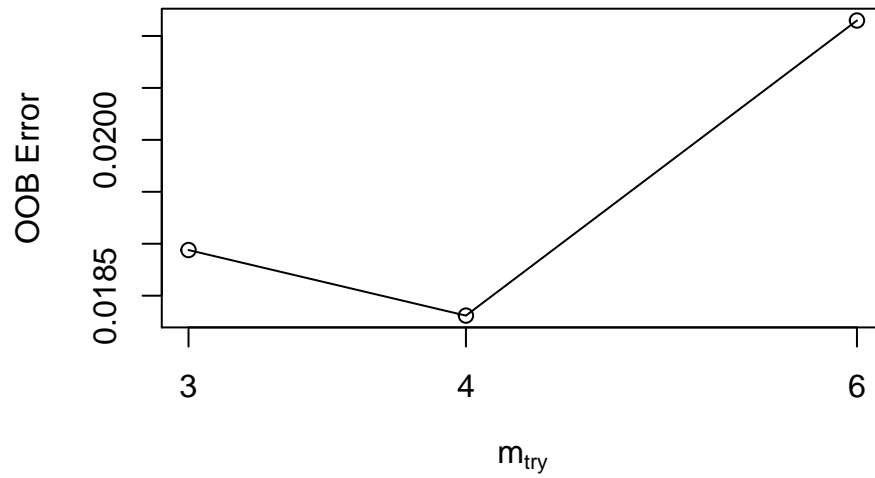
1. Find the best  $m_{try}$  (number of variables selected at each split)
2. According to plot of importance in the descending order, we select top 7 important features.

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

```

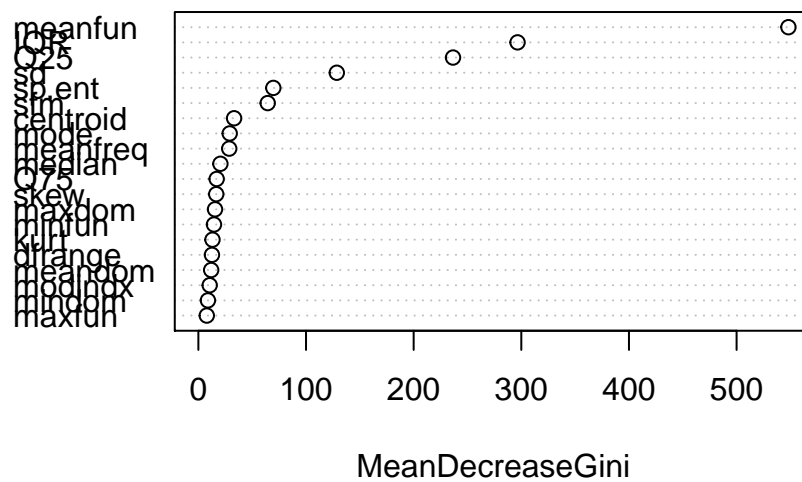
Searching right ...
mtry = 6    OOB error = 2.11%
-0.1551724 0.001

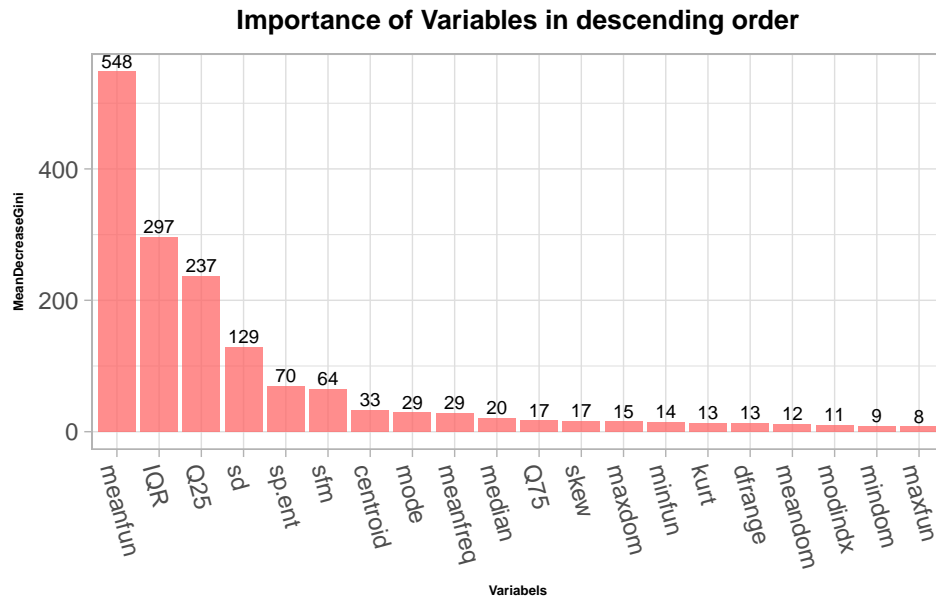
```



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

### trainingmodel





```
[1] "The selected features and the target variable are:"
[1] "sd"      "Q25"      "IQR"      "sp.ent"    "sfm"      "centroid"
[7] "meanfun" "label"
```

## IV. Models Building

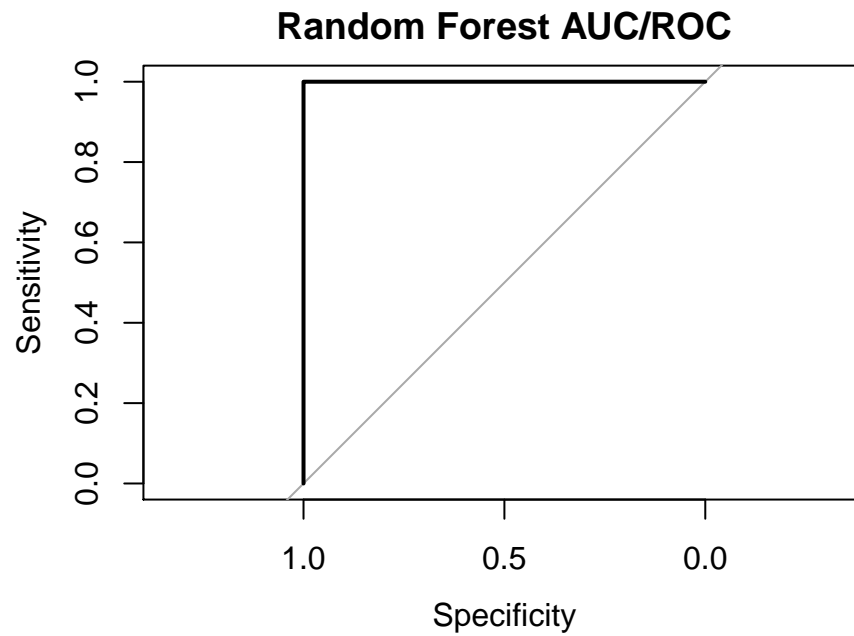
### 1. Random Forest Classification

- 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 the Random Forest model, the accuracy rate is:"
[1] "100 %"
```





#### Confusion Matrix and Statistics

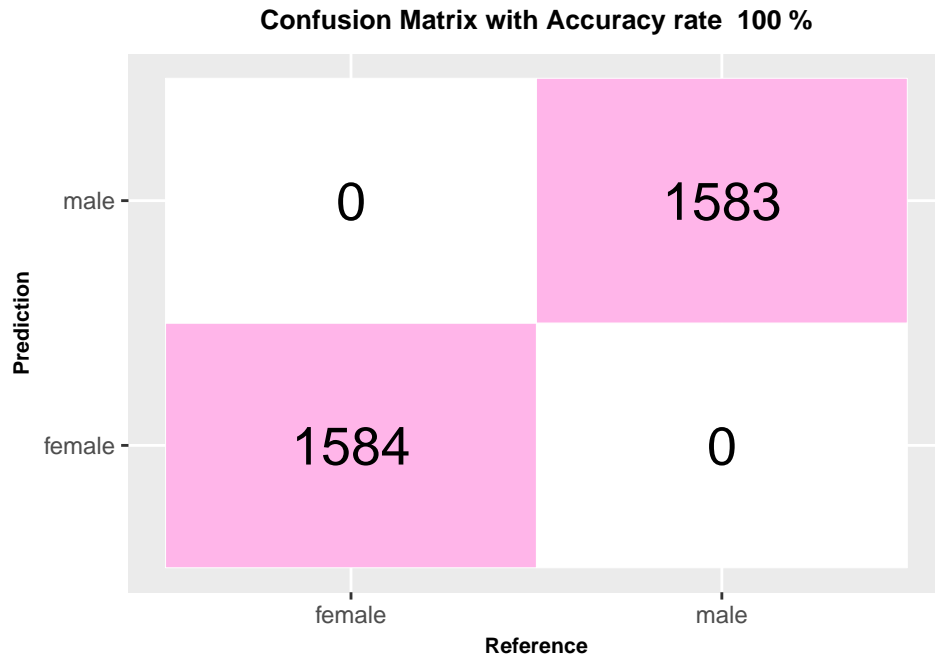
	Reference	
Prediction	female	male
female	1584	0
male	0	1583

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



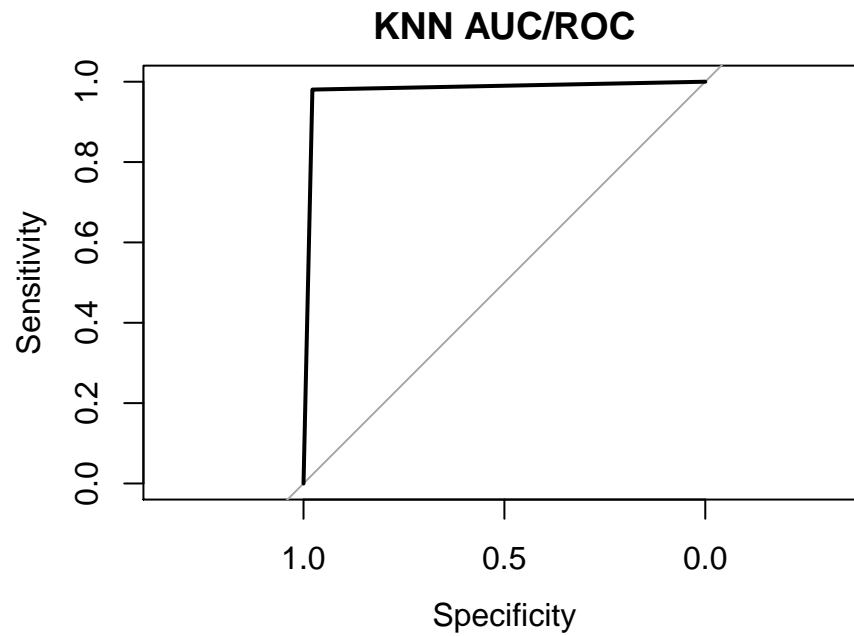
Based on the confusion matrix, there are 1584 samples are predicted as females and in the fact they are females; there are 1583 samples are predicted as males and in the fact they are males. Therefore, we can say that in the testing dataset, the prediction is 100% accurate.

The first segment of line in AUC is parallel to the y-axis and the angel of the left corner is 90 degree, which means that the area under curve is 100, and also means that this is a good model.

The accuracy 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.

## 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] "97.92 %"
```

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

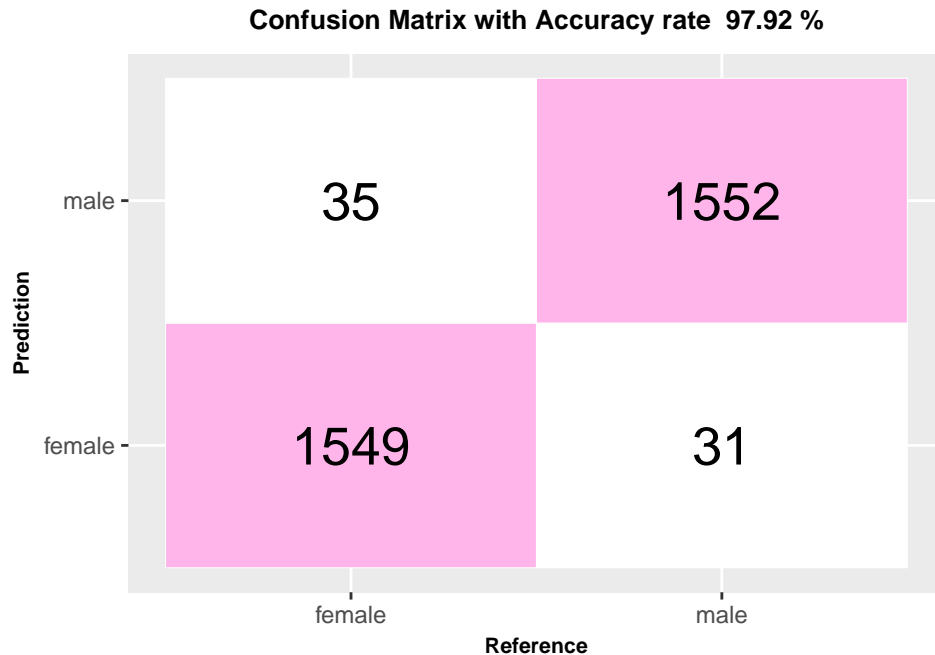
Prevalence : 0.5002

Detection Rate : 0.4891

Detection Prevalence : 0.4989

Balanced Accuracy : 0.9792

'Positive' Class : female



Based on the confusion matrix, there are 1549 samples are predicted correctly as females; however, there are 35 samples are predicted as male but actually they are female. There are 1552 samples are predicted correctly as males; however, there are 31 samples are predicted at female but they are in fact are male.

The accuracy is 97.92% 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
-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

Based on the summary table above, we find that the P-values of `sd`, `Q25` and `centroid` are larger than 0.05, which means those variables are not significant. Therefore, we need to remove those variables 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
-3.0790 -0.0358  0.0003  0.1105  4.2811

```

Coefficients:

```

              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -25.327     5.347  -4.737 2.17e-06 ***
IQR           59.100     4.329  13.651 < 2e-16 ***
sp.ent       53.995     6.357   8.494 < 2e-16 ***
sfm          -15.092     1.502 -10.049 < 2e-16 ***
meanfun     -160.073     8.122 -19.710 < 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: 613.01 on 3163 degrees of freedom
AIC: 623.01

```

Number of Fisher Scoring iterations: 8

From the second summary, we find that:

1. all the variables in the table above are very significant 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 59.1 For every one unit increase in `sp.ent`, the log odds of gender as female increase by 53.995. For every one unit increase in `sfm`, the log odds of gender as female decrease 15.092 (more likely to be male). For every one unit increase in `meanfun`, the log odds of gender as female decrease 160.073 (more likely to be male).
3. The Residual deviance increase from 610.18 to 613.01, which means the model is fitting good.
4. The Deviance Residuals is symmetrically distributed at center, not skewed to the left or right.
5. the AIC dropped from 626.18 to 623.01, which means the this model is fitting good as well.

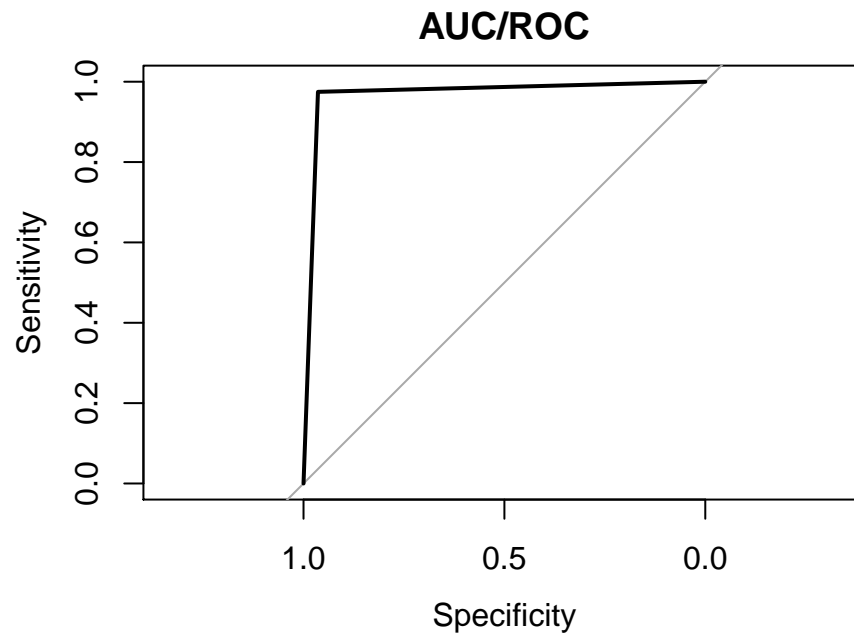
```

      actual
predictions female male
0      1527    40
1       57 1543

```

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

```
[1] "96.94 %"
```



#### Confusion Matrix and Statistics

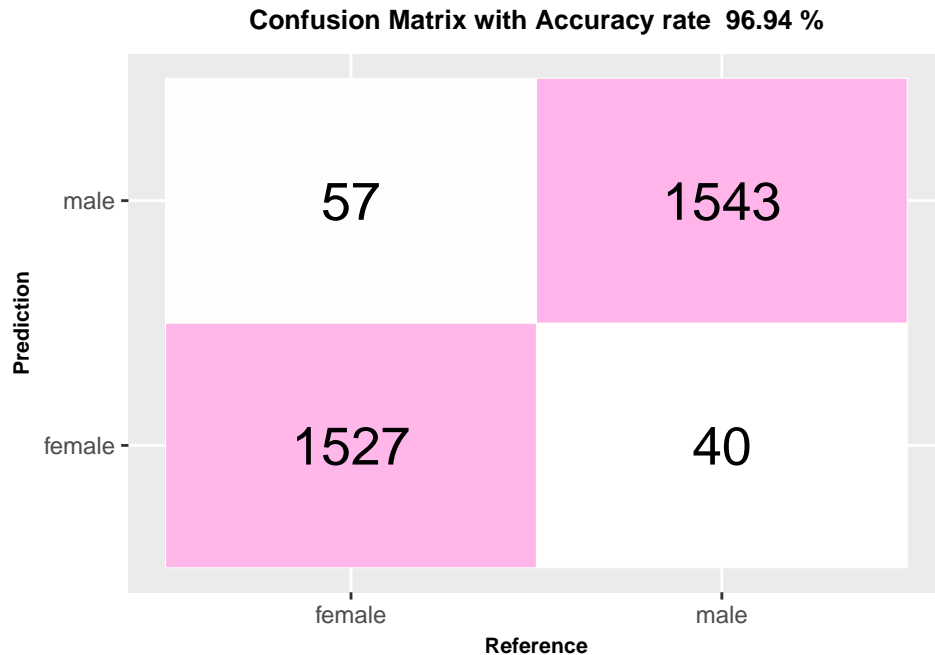
	Reference	
Prediction	female	male
female	1527	40
male	57	1543

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



Based on the confusion matrix, there are 1527 samples are predicted correctly as females; however, there are 57 samples are predicted as male but actually they are female. There are 1543 samples are predicted correctly as males; however, there are 40 samples are predicted at female but they are in fact are male.

The accuracy is 96.94% 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.

## V. Conclusion

- The accuracy rates of all three models are over 96%, they are also doing good on prediction. Among these three, the random forest classification is the best, since its accuracy reaches 100%. Therefore, we will use random forest model to do gender recognition.
- Gender can be recognized by voice. We will do a demo in presentation to show the gender recognition processing in class.
- 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?
- In the process of feature selection, there might be some better algorithms to do the dimension reduction such as principal component analysis (PCA) or independent component analysis (ICA).
- The 100% accurate rate in random forest is not true in real life. It is only suitable in the testing data set. If we add new sample, or make 10-fold cross validation, the accuracy rate will definitely not 100% and will be more 'accurate'.