남녀 목소리는 어떻게 다를까?

1. 윤대혁(19940524)
2. dbseogur2000@naver.com
3. 윤대혁(19940524)/우수경(19950722)

- 출처 : KBS '안녕하세요' 일부 캡쳐

youtube 영상 링크 : https://www.youtube.com/watch?v=_hfJHF7ORC0

위 사진에 나온 남자와 여자는 각각 목소리에 대한 고민을 가지고 있다.

여자는 너무 남자 같은 목소리를, 남자는 너무 여자 같은 목소리를 가지고 있는 것이다.

여기서 남자 같은 목소리와 여자 같은 목소리는 무엇을 말하는 것일까?

우리는 수화기너머로 들리는 목소리가 아는 사람이 아닐지라도, 성별정도는 쉽게 판단할 수 있다.

어쩌면 너무 당연하게 판단했기 때문에 굳이 고민해보지 않은 부분이기도 하다.

이를 판단하는데 가장 영향력 있는 요인은 무엇이 있을지,

과연 컴퓨터도 음성데이터를 분석하여 목소리의 성별을 분류할 수 있을지 궁금증이 생겼다.

따라서 본 데이터 분석을 통해 컴퓨터에 귀를 달아주려 한다.

기본적으로, 여자 목소리와 남자 목소리에는 주파수의 차이가 있다고 알려져 있다.

바꿔 말하면, 여자 목소리는 좀 더 고음이, 남자 목소리는 저음이 많다.

하지만 이 외에도 다른 구분 기준이 있을지 분석을 통해 알아보고자 한다.

분석에는 kaggle에서 제공하는 "Gender Recognition by Voice" 데이터를 사용하였다.

- ✓ 데이터 출처 : Kaggle open data set, 'Gender Recognition by Voice' by Kory Becker (https://www.kaggle.com/primaryobjects/voicegender/home)
- ✓ 남녀목소리 sample data를 R의 seewave, tuneR package를 사용하여 acoustic analysis[1]하여 전처리한 데이터
- ✓ 3168명의 목소리 데이터(남 : 1584명, 여 : 1584명)

변수 이름과 설명

1. meanfreq : 진동수(frequency)의 평균(단위 : khz)
2. sd : 진동수의 표준편차
3. median : 진동수의 중앙값(단위 : khz)
4. Q25 : 진동수의 제 1 사분위수(단위 : khz)
5. Q75 : 진동수의 제 3 사분위수(단위 : khz)
6. IQR : 진동수 사분위수 범위(= Q75 – Q25)
7. skew : 진동수 분포의 왜도
8. kurt : 진동수 분포의 첨도
9. sp.ent : spectral entropy
10. sfm : spectral flatness
11. mode : 진동수 최빈값
12. centroid : 진동수 중심값. Sum{frequency*(i 번째 주파수의 상대적 진폭)}으로 구해진다.
13. meanfun : acoustic signal 에 의해 측정된 기본(fundamental)진동수[2]의 평균
14. minfun : acoustic signal 에 의해 측정된 기본(fundamental)진동수의 최솟값
15. maxfun : acoustic signal 에 의해 측정된 기본(fundamental)진동수의 최댓값
16. meandom : acoustic signal 에 의해 측정된 지배(dominant)진동수[3]의 평균
17. mindom : acoustic signal 에 의해 측정된 지배(dominant)진동수의 최솟값
18. maxdom : acoustic signal 에 의해 측정된 지배(dominant)진동수의 최댓값
19. dfrange : 지배진동수의 범위(= maxdom – mindom)
20. modindx : 목소리의 억양을 나타내는 지표.
21. label : 성별(남자/여자)

[1] acoustic analysis: 발성의 질을 진동수, 강도, 시간적 측면에서 분석한 것.

[2] 기본 진동수 : 사람의 목소리는 여러 주파수의 복합음으로 되어있는데, 이들을 모두 분해하여 스펙트럼으로 나타낼 때 가장 낮은 위치에 있는 주파수

[3] 지배 진동수 : 주파수 스펙트럼에서 가장 빈번하게 등장하는 진동수

0. 데이터 샘플

다음은 주어진 데이터로부터 랜덤하게 5 개를 뽑은 것이다. 데이터의 대략적인 생김새를 알 수 있다.

데이터를 보면 목소리 주파수들의 분포를 알려주는 측도들이 변수로 되어있다.

Meanfreq	Sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	mode
0.18094	0.05857	0.18631	0.14706	0.22190	0.07484	1.22387	4.54343	0.93024	0.47638	0.18788
0.19880	0.03738	0.20777	0.18393	0.21844	0.03451	2.79094	12.1811	0.84655	0.25599	0.21275
0.21192	0.05064	0.21184	0.19710	0.24905	0.05194	2.20193	8.60356	0.87260	0.28467	0.20226
0.19176	0.06726	0.21350	0.18086	0.24059	0.05972	2.52209	10.4659	0.92576	0.56021	0.24998
0.15853	0.05921	0.13728	0.19710	0.21547	0.10589	2.94200	15.31788	0.92005	0.45201	0.11063

centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx	Label
0.18094	0.11680	0.04918	0.27745	0.73289	0.02343	3.67968	3.65625	0.14102	male
0.19880	0.19188	0.04907	0.22857	0.39843	0.16406	3.91406	3.75	0.11226	female
0.21192	0.18926	0.04705	0.27428	1.59635	0.02343	12.0234	12	0.10095	female
0.19176	0.18717	0.01656	0.26666	0.35468	0.00781	4.85937	4.85156	0.07497	female
0.15853	0.10952	0.04566	0.23809	0.74275	0.00488	3.99902	3.99414	0.26665	male

1. 남녀의 목소리는 어떠한 특징이 있을까?

남녀의 목소리가 어떻게 다른 지 알아보기 위해 주어진 데이터를 남자 데이터(=label 변수의 male)와

여자 데이터(label 변수의 female)로 나눈 뒤, 각 변수마다 상자그림(boxplot)과 히스토그램(histogram)을 그려

전체적인 분포를 눈으로 확인해보았다.

그 결과 남녀간 차이 정도에 따라, 뚜렷함, 애매함, 동일함 세 그룹으로 나눌 수 있었다.

(뚜렷함 : 남녀간 차이가 뚜렷한 경우, 애매함 : 남녀간 차이가 애매한 경우, 동일함 : 남녀간 분포가 거의 비슷한 경우)

표로 정리하면 다음과 같다.

뚜렷함	IQR, meanfun, Q25, sd, sp.ent, sfm (6)
애매함	meanfreq, centroid, median, dfrange, maxdom, meandom, mode (7)
동일함	skew, kurt, maxfun, modindx, Q75, mindom, minfun (7)

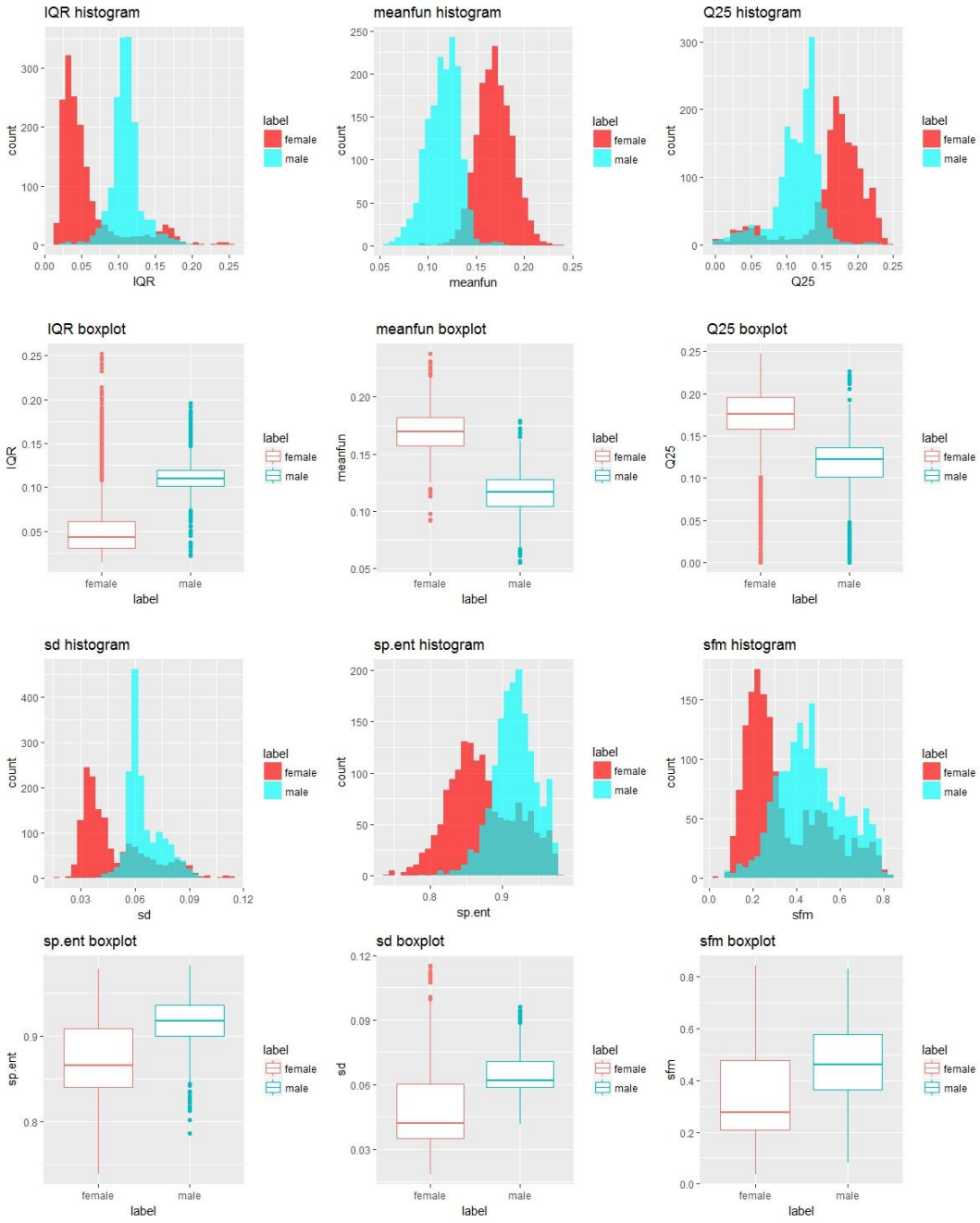
첫번째 6 개의 변수는 남녀간 차이가 눈에 띄게 나타났다.

두번째 7 개의 변수는 남녀간 차이가 나타난다고 보기 애매했고

마지막 7 개의 변수는 남녀간 차이가 나타나지 않는다고 볼 수 있었다.

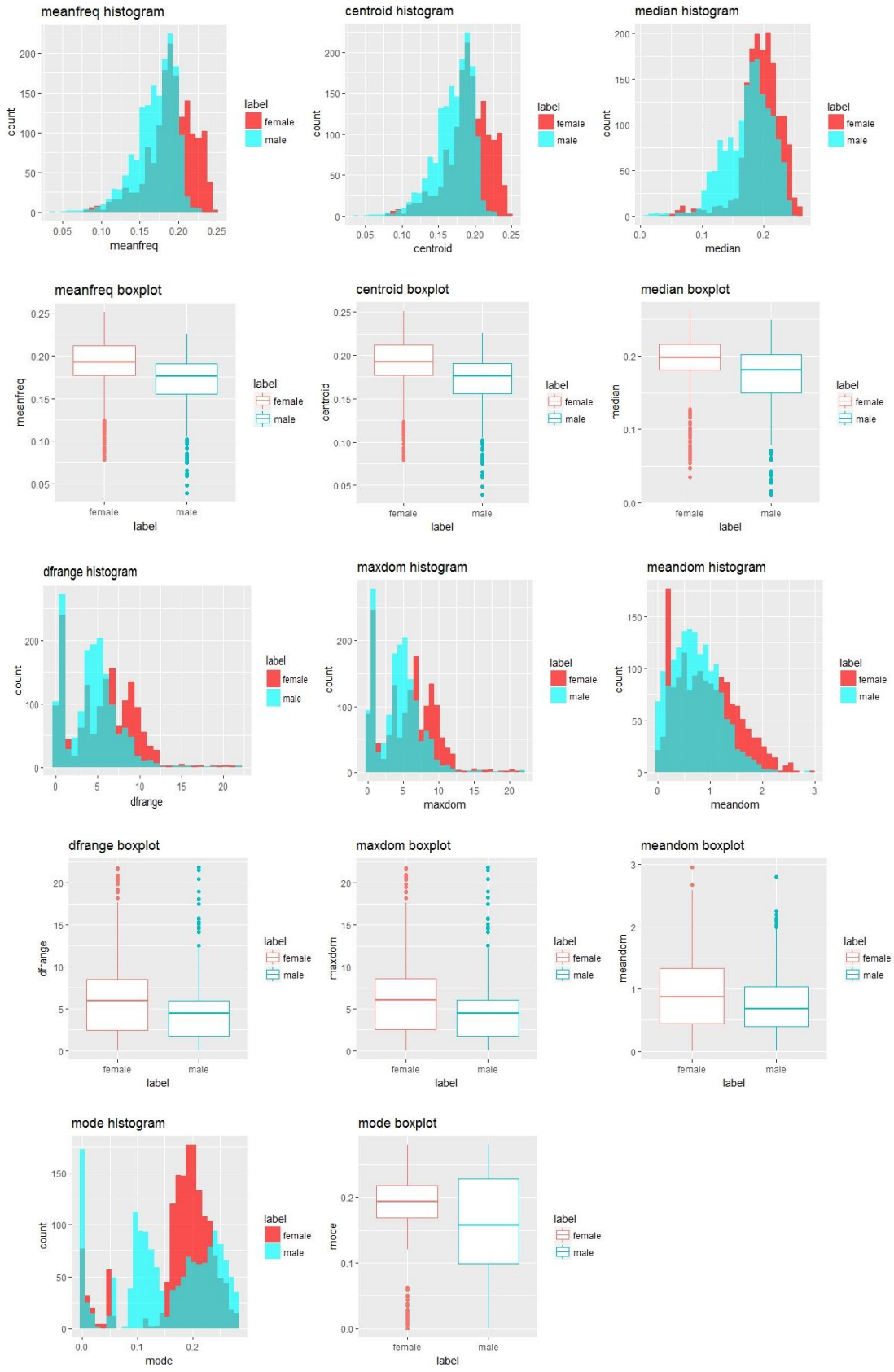
이제 위와 같이 분류한 근거로써, 그룹별, 변수별로 히스토그램 및 상자그림을 살펴보자.

1) 뚜렷함



위에 제시된 6 개의 변수 IQR, meanfun, Q25, sd, sp.ent, sfm 은 남녀간 분포의 차이가 눈에 띄게 나타났음을 확인할 수 있었다. 이들 변수는 남녀 목소리를 구분하는데 있어서 매우 중요한 역할을 할 수 있을 것으로 예상할 수 있다.

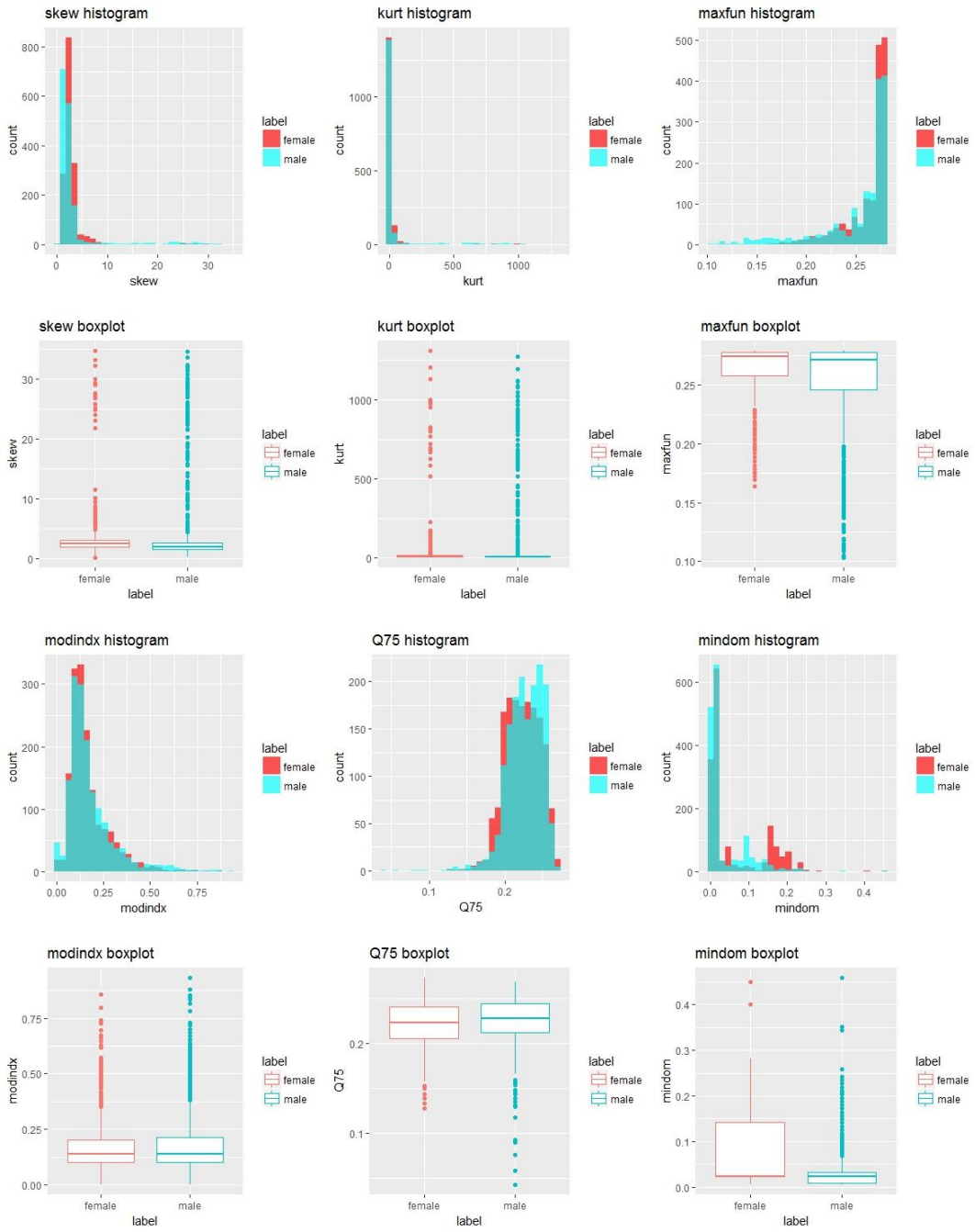
2) 애매함

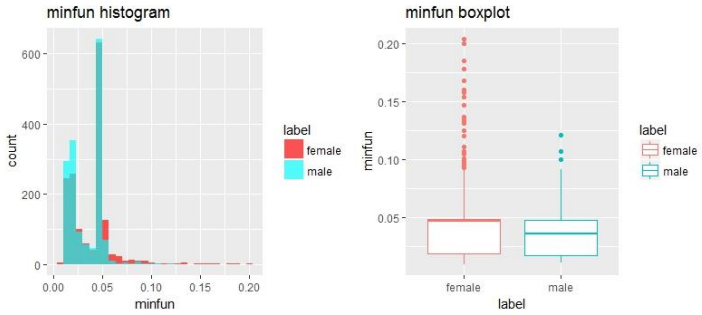

meanfreq histogram
count
meanfreq
centroid histogram
centroid
median histogram
median
label
female
male
meanfreq boxplot
centroid boxplot
median boxplot
dfrange histogram
dfrange
maxdom histogram
maxdom
meandom histogram
meandom
dfrange boxplot
maxdom boxplot
meandom boxplot
mode histogram
mode
mode boxplot

위에 제시된 7 개의 변수 중 meanfreq, centroid, median, dfrange, maxdom, meandom 6 개의 경우는 여자가 남자보다 그 값이 더 큰 경향이 있고 mode 의 경우는 여자의 경우 특정 구간에 남자보다 더 집중해 있는 경향이 보이기 때문에 남녀간의 차이가 조금 나타난다고 예상할 수 있었다.

그러나 이들은 사람에 따라 큰 차이가 있지 않다고 판단할 수 있을 것 같다.

3) 동일함





위에 제시된 7 개의 변수 skew, kurt, maxfun, modindx, Q75, mindom, minfun 은 남녀 간에 큰 차이를 보이지 않았다. 이들은 남녀 성별을 구분하는데 큰 영향을 끼치지 않을 것이라고 예상할 수 있다.

2. 남녀를 구분하기 위한 목소리의 기준은 어떻게 생각해야 좋을까?

상자그림과 히스토그램은 두 집단의 분포를 시각적으로 확인할 수 있는 장점이 있지만, 사람마다 다르게 판단할 수 있는 부분도 있고, 실제로 두 집단의 분포가 다르다고 확신할 수는 없다. 이를 위해 통계적 가설검정을 실시한다.
두 집단을 비교하기 위한 대표적인 가설검정은 이표본 t 검정(two sample t-test)이 있는데, 이를 위해서는 두 집단의 분포가 정규분포를 따라야 한다는 가정이 필요하다. 하지만 히스토그램을 확인해보면 모든 변수가 정규분포를 따른다고 보기에는 무리가 있다. 실제로 그런지 확인해보기 위해 정규성 검정 중 Shapiro-wilk test 를 실시하여 p-value 를 확인해보았다.

<Shapiro-wilk test p-value>

변수	남자	여자	변수	남자	여자
Meanfreq	3.356687e-24	1.454154e-20	**Median**	2.789751e-19	2.13413e-30
Sd	1.144471e-28	2.411574e-33	**Centroid**	3.356687e-24	1.454154e-20
Q25	7.549999e-29	4.967395e-38	**Meanfun**	2.123894e-10	0.0001443098
Q75	4.686175e-29	1.67468e-12	**Minfun**	9.314719e-39	3.25596e-42
IQR	9.952656e-25	1.150833e-44	**Maxfun**	6.592798e-46	1.809254e-45
Skew	5.114974e-58	3.381002e-59	**Meandom**	6.857537e-16	1.073065e-19
Kurt	3.385069e-61	3.657622e-64	**Mindom**	3.256263e-48	9.722639e-45
Sp.ent	1.429096e-11	4.56461e-14	**Maxdom**	6.34671e-27	8.630046e-24
Sfm	2.599919e-11	2.279641e-31	**Dfrange**	4.911282e-27	8.845618e-24
mode	3.331737e-26	5.210108e-40	**modindx**	1.038452e-39	2.109361e-37

귀무가설은 "H_0 : 집단이 정규분포를 따른다."인데, 위와 같이 모든 변수에 대해 p-value 가 매우 낮아 남녀 집단의 분포는 정규분포를 따르지 않는다고 주장할 수 있다.

각 집단의 변수 분포가 정규분포를 따르지 않을 때, 두 집단을 비교하는 통계적 검정으로 비모수 검정인 Wilcoxon rank sum test(Mann–Whitney U test)가 있다. 이를 실시하여 두 집단이 실제로 차이가 있다고 주장할 수 있는지 알아보았다.

< Wilcoxon rank sum test p-value>

변수	p-value	변수	p-value	변수	p-value
Meanfreq	4.609958e-91	Sfm	2.670191e-106	Meandom	4.646993e-20
Sd	1.095488e-169	Skew	1.753031e-48	Mindom	1.341577e-20
Median	3.269232e-61	Kurt	7.757113e-29	Maxdom	2.201108e-30
Q25	3.012436e-268	Mode	2.379333e-15	Dfrange	8.978694e-29
Q75	9.163526e-07	Meanfun	0	Modindx	0.7476314
IQR	7.89069e-289	Minfun	3.767347e-14	centroid	4.609958e-91
Sp.ent	7.644954e-160	Maxfun	2.150572e-12		

P-value 를 확인해보면 modindx 의 경우 매우 높은 값을 가져 이는 두 집단을 구별하기에 적절하지 않다고 생각할 수 있다. 이는 히스토그램과 상자그림을 통해 이미 예상한 결과이다.

통계적 검정으로 modindx 를 제외한 모든 변수가 남녀집단에서 차이가 있다는 것이 통계적으로 유의하다는 것을 확인할 수 있었다. 하지만 두 집단 간 차이의 정도는 변수마다 다르다. 어떤 변수는 그 차이가 근소할 수 있고, 어떤 변수는 그 차이가 매우 클 수 있다. 당연히 두 집단을 구분하는데 그 차이가 클수록 눈여겨봐야 할 변수라고 할 수 있겠다.

그렇다면 어떤 변수가 남녀 간의 차이를 만들어 내는지 알 수 있을까? 명확한 수치적 기준이 필요하다. 먼저 각 집단의 기본적인 대푯값인 평균의 차이를 확인해보았다. 이 때, 각 변수는 측정 단위가 다르므로 자료를 해당 변수의 표준편차로 나눈 뒤 평균을 구해 차이를 계산하여 변수끼리 비교할 수 있도록 하였다.

<변수에 따른 남녀 평균차이>

변수	차이	변수	차이	변수	차이
Meanfreq	0.1309796	Sfm	1.217682	Meandom	0.02294481
Sd	4.212953	Skew	0.3237837	Mindom	0.07416286
Median	1.406733	Kurt	0.0423307	Maxdom	0.03565802
Q25	0.2889374	Mode	0.8485932	Dfrange	0.03158902
Q75	0.2414981	Meanfun	2.44996	Modindx	0.2165606
IQR	4.079325	Minfun	0.3649685	Centroid	0.1309796
Sp.ent	13.23205	Maxfun	5.205352		

이를 이용하면 5 개의 변수 sd, meanfun, maxfun, IQR, sp.ent 가 집단간 차이가 다른 변수에 비해 비교적 크다는 것을 확인할 수 있다.

그러나 흥미로운 점은 maxfun 과 Q25 이다. 히스토그램과 상자그림을 확인해보면 maxfun 은 두 집단 분포의 차이가 거의 없다고 할 수 있었고 Q25 는 이에 비해 큰 차이가 나타났다. 그러나 평균 차이는 Q25 는 거의 없고 maxfun 은 차이가 크게 나타났다. 왜 이런 현상이 나타났을까?

이에 대한 해답은 의외로 간단하다! 사람의 목소리는 여러 주파수의 복합음인데 이 복합음 각 주파수의 진동수에 대해 계산한 제 1 사분위수 Q25 는 차이가 작아 보이고 복합음을 분해하여 얻은 순음 fundamental frequency 에서 최댓값을 계산한 maxfun 은 당연히 그 차이가 커 보인다. 이러한 점을 평균이 반영하지 못하고 있다.
따라서 우리는 좀 더 좋은 기준을 가지고 영향력 있는 변수를 찾아낼 필요가 있다.

그렇다면 로지스틱 회귀분석을 이용해보면 어떨까? 주어진 3168 개의 데이터를 랜덤하게 7:3 의 비율로 훈련데이터(training)와 시험데이터(test)로 나눈 뒤 훈련데이터에 대해 반응변수를 성별(label)로 하고 20 개의 설명변수에 대해 각 변수마다 로지스틱 회귀모형을 적합 시켜 시험데이터를 이용하여 분류 예측을 해보도록 한다. 정확도(accuracy)[4]가 높을수록 해당 설명변수는 두 집단을 더 잘 구별해준다고 생각할 수 있다.

<변수의 분류 정확도>

변수	정확도	카파(kappa)[5]	변수	정확도	카파(kappa)
Meanfreq	0.6425	0.2845	Sfm	0.6677	0.3357
Sd	0.7834	0.5677	Meanfun	0.9558	0.9117
Q25	0.8528	0.7057	Minfun	0.5457	0.0911
Q75	0.5163	0.0333	Maxfun	0.5468	0.0892
IQR	0.8864	0.7731	Meandom	0.572	0.1452
Skew	0.4006	-0.2021	Mindom	0.5521	0.1087
Kurt	0.5121	0.0151	Maxdom	0.5868	0.1738
Median	0.6267	0.2529	Dfrange	0.5846	0.1695
Mode	0.6446	0.2875	Centroid	0.6425	0.2845
Sp.ent	0.7445	0.4897	Modindx	0.5037	0.0046

[4] 정확도 : 전체 데이터에 대하여 예측값과 실제값이 일치하는 경우에 대한 비율

[5] 카파(kappa) : 우연히 정확히 예측할 확률값을 이용해서 조정된 정확도

정확도가 50%에 가깝다면, 두 집단을 구별하는데 큰 영향을 미치지 않는다고 볼 수 있다.
Sd, Q25, IQR, sp.ent, meanfun 5 개의 변수는 정확도가 모두 70%이상으로 남녀를 구분해내고 있다.
이들은 카파(kappa)통계량도 다른 변수들에 비해 대체로 높다. 이들이 남녀를 구분하는데 매우 유용한 변수라고 생각할 수 있다. 특히 재미있는 점은 남녀 목소리의 전반적인 높고 낮음을 나타내는 meanfun 은 96% 정확도로 남녀를 구분하고 있는데, 직관적으로 당연하다!

한편, maxfun 의 경우 앞서 평균차이를 이용했을 때와 달리, 정확도 및 카파통계량을 기준으로 했을 때 남녀차이를 구별하는데 큰 영향을 주는 변수가 아님을 알 수 있다.

3. 남녀 목소리 구분 모형(Full model)

지금까지 데이터를 탐색해본 결과, 5 개의 변수 sd, sp.ent, IQR, meanfun, Q25 가 남녀를 구별하는데 가장 중요한 변수임을 확인할 수 있었다. 특히 직관적으로 남녀 목소리의 전체적인 높고 낮음을 나타내는 meanfun 은 남녀를 구별하는데 매우 중요한 변수였다.

이렇게 단순한 구별보다는 이들을 종합적으로 고려한 통계적인 모형을 만들어보면 어떨까?
그렇다면 조금 더 높은 정확도를 가지고 남녀를 구별할 수 있을까?
먼저 modindx 를 제외한 모든 변수는 통계적 검정으로 남녀간 차이가 있다고 말할 수 있었다. 비록 어떤 변수는 차이가 크지 않더라도 남녀를 구별하는데 영향을 줄 수 있기 때문에 먼저 modindx 를 제외한 모든 변수를 고려하도록 하자.

적합 과정은 다음과 같다.
주어진 3168 개의 데이터를 6:2:2 의 비율로 랜덤하게 훈련(training)/검증(validation)/시험(test)데이터로 나눈다.

```
> length(training_index)
[1] 1900
> length(validate_index)
[1] 633
> length(test_index)
[1] 635
```

고려하고자 하는 모형은 분류 문제에 사용되는 대표적인 5 가지 모형인 logistic regression model, decision tree, random forest, support vector machine, gradient boosting model 이다.

훈련 데이터를 이용하여 각 모형을 적합 시키고, 검증 데이터를 이용하여 예측을 시도하여 모형의 성능을 평가한다.

성능 평가의 기준으로 사용되는 것은 AUC, binomial deviance[6], ROC curve, accuracy, kappa 이다. 이들을 종합적으로 고려하여 가장 좋은 모형을 택한다.

가장 좋은 모형을 이용하여 test 데이터로 예측을 시도한다. 이로부터 confusion matrix 를 계산하여 예측력을 평가해본다.

1) Logistic regression model

Modindx 를 제외한 모든 변수를 고려하여 로지스틱 회귀모형을 적합 시키는 경우 다음과 같이 설명변수간 상관계수가 높은 변수들이 존재하여 회귀계수가 추정되지 않는 경우가 발생한다.

```
> cor(training_voice$meanfreq,training_voice$centroid)
[1] 1
> cor(training_voice$IQR,training_voice$Q25)
[1] -0.8695692
> cor(training_voice$dfrange,training_voice$maxdom)
[1] 0.999843
```

그러므로 centroid 와 Q25, maxdom 은 제외하도록 하자. 특히 실제로 데이터를 확인해본 결과 centroid 와 meanfreq 는 같은 값을 갖는 변수였고 의미적으로 와 닿는 변수는 meanfreq 이기 때문에 centroid 는 제거하고 Q25 보다는 IQR 이 Q75-Q25 로 계산된 값이므로 정보를 함축적으로 갖고 있다고 생각하여 Q25 를 제거하고 dfrange 는 maxdom-mindom 으로 계산된 값이므로 마찬가지 이유로 maxdom 을 제거한다.
모형간 비교를 위해 다른 모형에서도 마찬가지로 이들을 제거하고 고려하도록 하자.

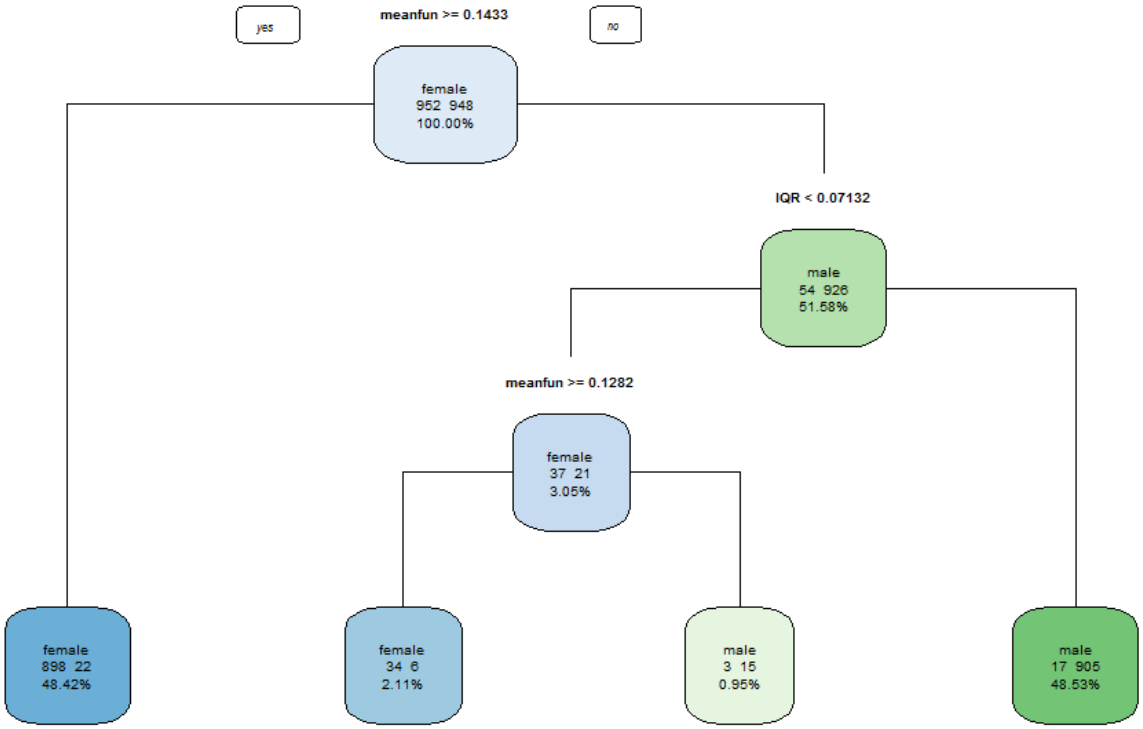
로지스틱 모형을 적합 시킨 뒤, 검증데이터를 이용하여 예측을 시도한 결과 confusion matrix 는 다음과 같았다.

```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    306    7
    male       10  310

               Accuracy : 0.9731
```

[6] binomial deviance : 모형의 정확도 지표의 일종으로, 값이 작을수록 모형이 정확하다. R 에서 제공되는 함수가 없기 때문에 직접 함수를 만들어 계산하였다.

2) Decision tree

의사결정나무 모형을 적합 시킨 결과 의사결정나무는 다음과 같았다.



이 모형을 이용하여 계산된 confusion matrix 는 다음과 같았다.

```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    301   11
    male       15  306

               Accuracy : 0.9589
```

3) Random forest

이 모형을 이용하여 검증데이터에 대해 예측을 시도하여 confusion matrix 를 계산하면 다음과 같았다.

```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    307    7
    male        9  310

               Accuracy : 0.9747
```

4) Support vector machine

Support vector machine 을 적합 시켜 검증데이터에 대해 예측을 시도한 결과 confusion matrix 는 다음과 같았다.

```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    310    6
    male        6  311

               Accuracy : 0.981
```

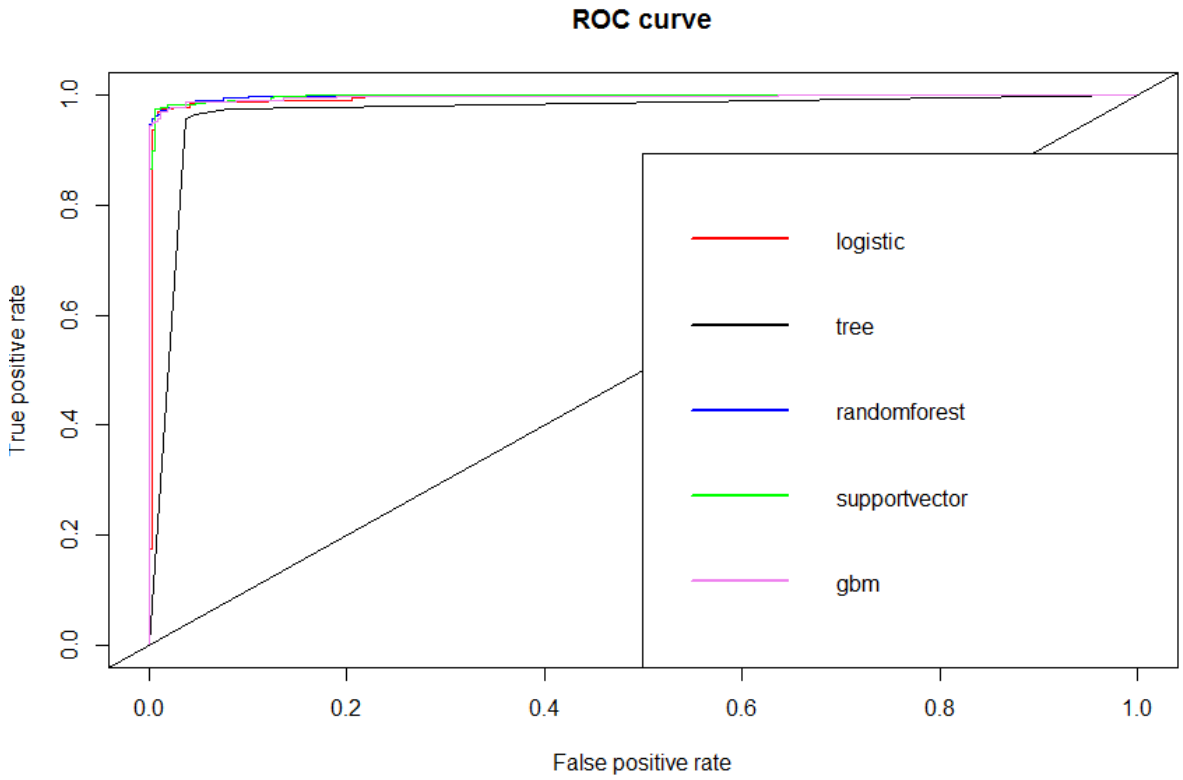
5) Gradient boosting model

Gradient boosting model 을 적합 시킨 뒤 검증데이터에 대해 예측을 시도하여 confusion matrix 를 계산하면 다음과 같았다.

```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    305    7
    male       11  310

               Accuracy : 0.9716
```

6) 모형 비교

먼저 5 가지 모형의 ROC curve 를 그려보면 다음과 같았다. 좌상단으로 curve 가 올라갈수록 모형의 예측력이 좋다고 할 수 있다. 어느 하나의 모형을 선택할 수 없도록 모든 모형의 예측력이 대체로 좋았다.



다음은 AUC, binomial deviance, accuracy, kappa 통계량을 계산하여 표로 정리한 결과이다.

AUC 는 높을수록, binomial deviance 는 낮을수록, accuracy 는 높을수록 kappa 는 높을수록 좋은 모형이라고 말할 수 있다. 이들을 종합적으로 고려해본 결과 모든 모형이 나쁘지 않았지만 random forest 모형이 가장 좋다고 말할 수 있었다.

	model_name	auc_model	binomial_model	accuracy_model	kappa_model
1	logistic	0.9938905	3518.438	0.9731438	0.9463
2	tree	0.9672314	5820.999	0.9589258	0.9178
3	randomforest	0.9976740	3628.961	0.9747235	0.9494
4	supportvector	0.9974444	5949.942	0.9810427	0.9621
5	gbm	0.9959070	5784.158	0.9715640	0.9431

7) 최종모형의 일반화 능력 평가

최종적으로 선택한 random forest 모형을 이용하여 시험데이터에 대해 예측을 시도한 결과 confusion matrix 는 다음과 같았다. 참고로 계산된 binomial deviance 는 3905.269, AUC 는 0.9968305, kappa 는 0.9339 였다.

```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    306   11
    male       10  308

               Accuracy : 0.9669
```

정확도 97%로 시험데이터를 예측하여 모형의 성능이 매우 좋음을 확인할 수 있다.

하지만 고려한 설명변수가 modindx,Q25,maxdom,centroid 를 제외한 16 개나 되어 모형이 매우 복잡하다는 점에서 아쉽다. 그렇다면 앞에서 선택한 남녀를 구분하는데 중요한 5 개의 변수만을 고려한 모형을 적합 시켜보면 어떨까?

4. 남녀 목소리 구분 모형(reduced model)

정확한 비교를 위해 동일한 데이터를 사용하고 마찬가지로 5 가지의 모형을 고려하였다.

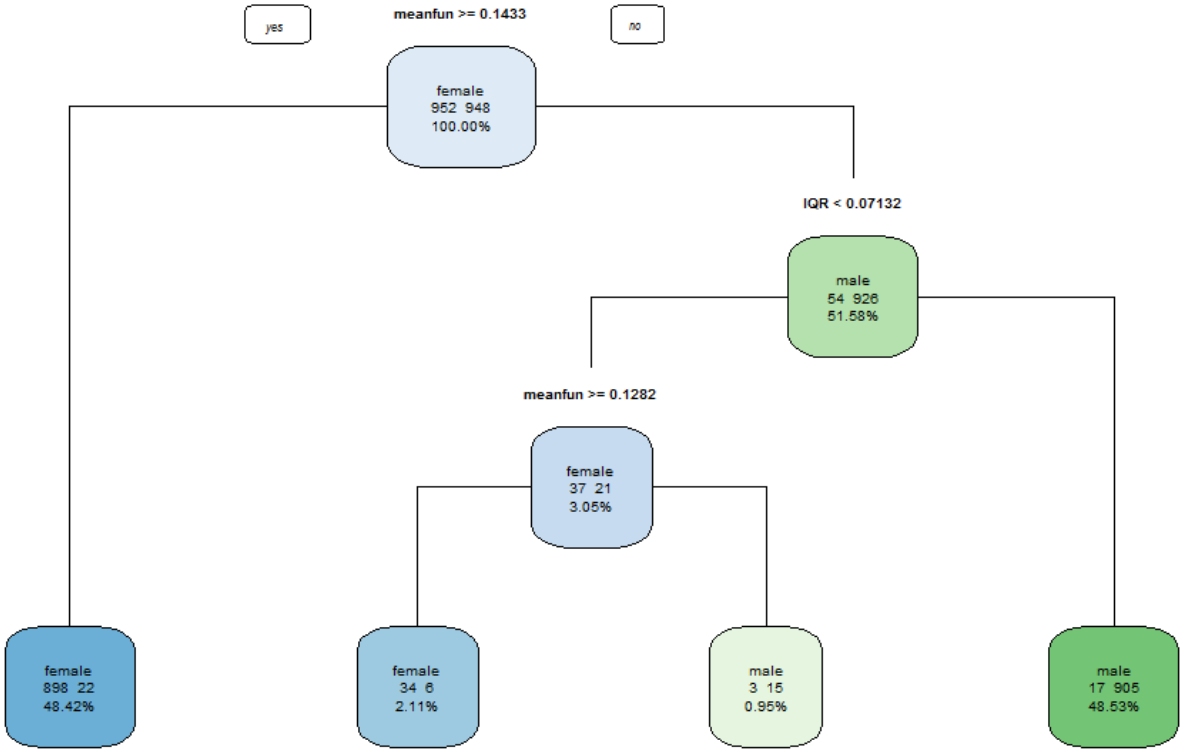
1) Logistic regression model

5 개의 변수(meanfreq, IQR, Q25, sd, sp.ent)를 사용하여 로지스틱 회귀모형을 적합 시켜 검증데이터에 대해 confusion matrix 를 계산하면 다음과 같았다.

```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    304    7
    male       12  310

               Accuracy : 0.97
```

2) Decision tree

의사결정나무 모형을 적합 시켜 얻은 의사결정나무 그림은 다음과 같았다.



검증데이터로부터 계산된 confusion matrix 는 다음과 같았다.

Confusion Matrix and Statistics

Prediction	Reference female	Reference male
female	301	11
male	15	306

Accuracy : 0.9589

3) Random forest

랜덤포레스트 모형을 적합 시켜 검증데이터로부터 계산된 confusion matrix 는 다음과 같았다.

```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    303    6
    male       13  311

               Accuracy : 0.97
```

4) Support vector machine

Support vector machine 을 적합 시켜 검증데이터로부터 계산된 confusion matrix 는 다음과 같았다.

```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    309    5
    male        7  312

               Accuracy : 0.981
```

5) Gradient boosting model

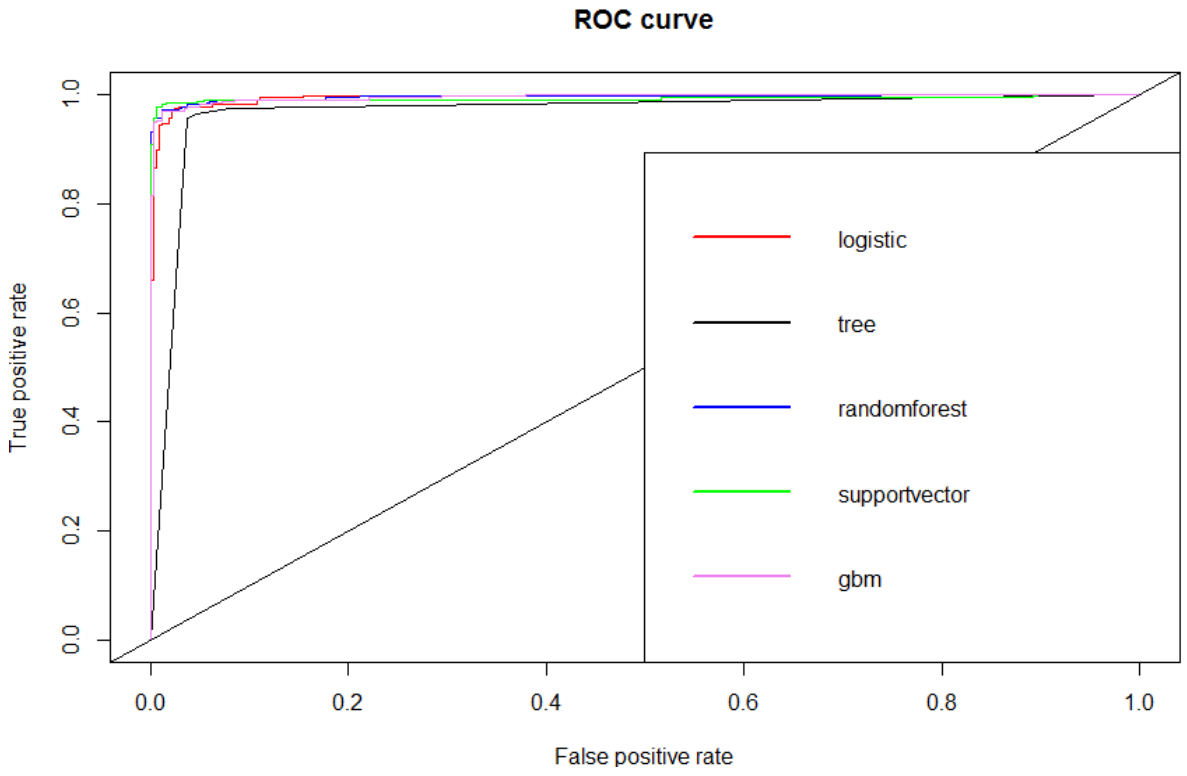
Gradient boosting model 을 적합 시켜 검증데이터로부터 계산된 confusion matrix 는 다음과 같았다.

```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    301    7
    male       15  310

               Accuracy : 0.9652
```

6) 모형 비교

먼저 5 가지 모형의 ROC curve 를 그려보면 다음과 같았다.

모든 모형이 비슷하게 좌상단에 위치하여 예측력이 훌륭하다고 말할 수 있어서 어느 하나의 모형이 좋다고 말하기 어렵다.



AUC, binomial deviance, accuracy, kappa 통계량을 계산하여 표로 정리하면 다음과 같았다.

	model_name	auc_model	binomial_model	accuracy_model	kappa_model
1	logistic	0.9943298	4457.883	0.9699842	0.9400
2	tree	0.9672314	5820.999	0.9589258	0.9178
3	randomforest	0.9952681	2800.040	0.9699842	0.9400
4	supportvector	0.9920337	5820.999	0.9810427	0.9621
5	gbm	0.9953280	5820.999	0.9652449	0.9305

4 가지 기준을 종합적으로 고려해 보았을 때 모든 모형이 훌륭하지만,

random forest model 이 가장 좋다고 말할 수 있다.

7) 최종모형의 일반화 능력 평가

적합 시킨 random forest 모델을 이용하여 시험데이터로부터 계산된 confusion matrix 는 다음과 같다.

참고로 계산된 AUC 는 0.9945389, binomial deviance 는 2634.256, kappa 는 0.9276 이었다.

```
Confusion Matrix and Statistics

          Reference
Prediction female male
    female    304   11
    male       12  308

               Accuracy : 0.9638
```

설명변수를 11 개나 줄였는데 16 개 사용한 모형(약 97%)과 비슷한 수준의 정확도 약 96%로 시험데이터를 예측하였다. 그 외 다른 예측력 지표 AUC, binomial deviance, kappa 도 떨어지지 않았다.

그렇다면 훨씬 간단한 이 모형을 최종적으로 선택할 수 있겠다.