Gender Recognition by Voice

DS 502 Final Project

Group Members:

Fangzheng Sun

Huimin Ren

Shanhao Li

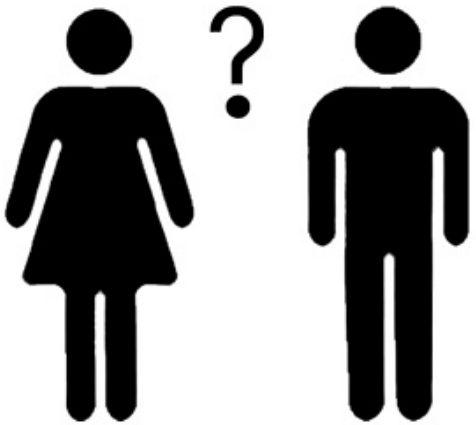
Sahil Shahani

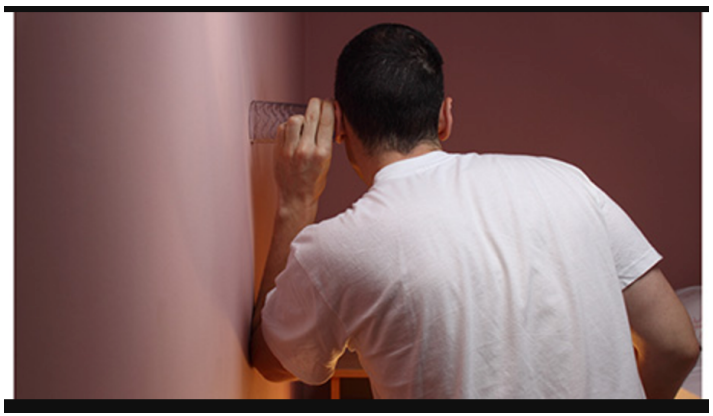
Yun Yue

Worcester Polytechnic Institute

April, 2016

Gender Recognition by Voice





Used for validating financial transaction, marketing and sales, government agencies

https://bkgyan.files.wordpress.com/2012/11/male-female.jpg
https://www.insidescience.org/sites/default/files/hear-top.jpg

Kaggle Voice Data

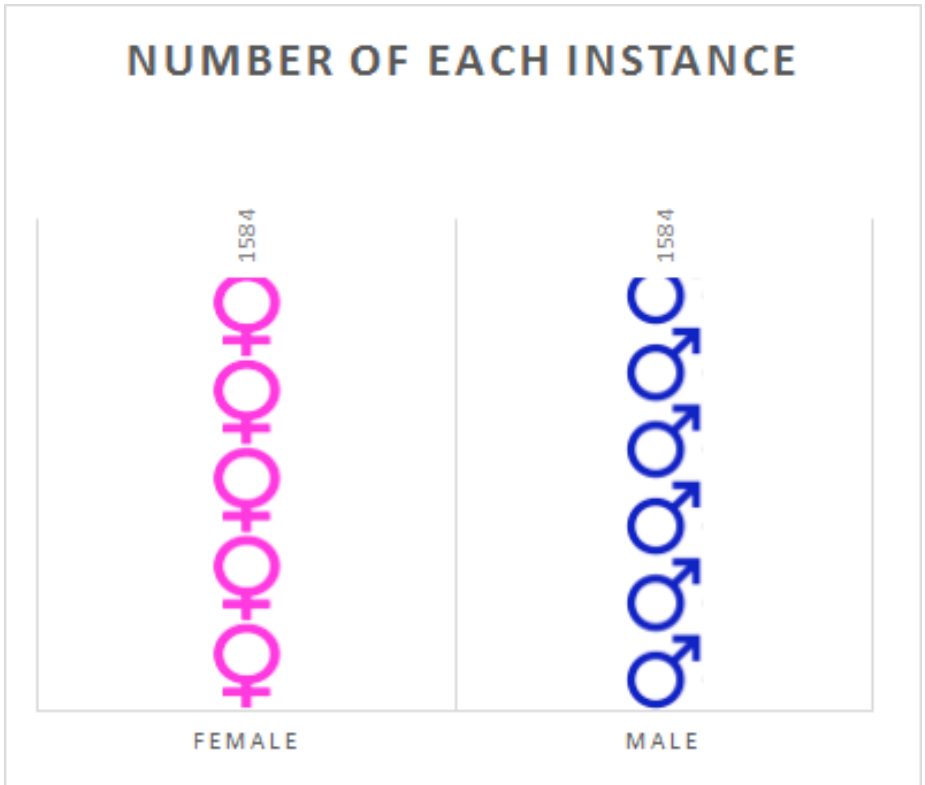
- Voice samples (3,168 records)

Kaggle Voice Data

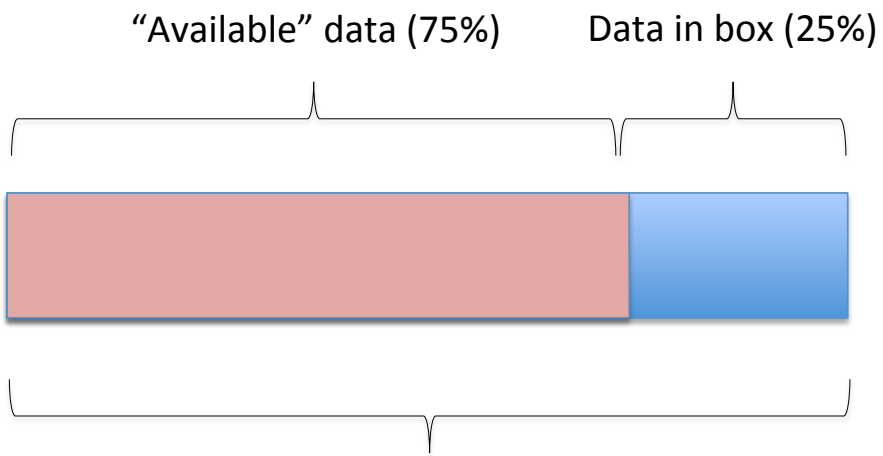
- Voice samples (3,168 records)
- 20 acoustic properties and 1 true value (male or female)

  meanfreq, sd, median, Q25, Q75, IQR, skew, kurt, sp.ent, sfm, mode, centroid, peakf, meanfun, minfun, maxfun, meandom, mindom, maxdom, dfrange, modindx, label

Overview of Data: Number of Each Instance

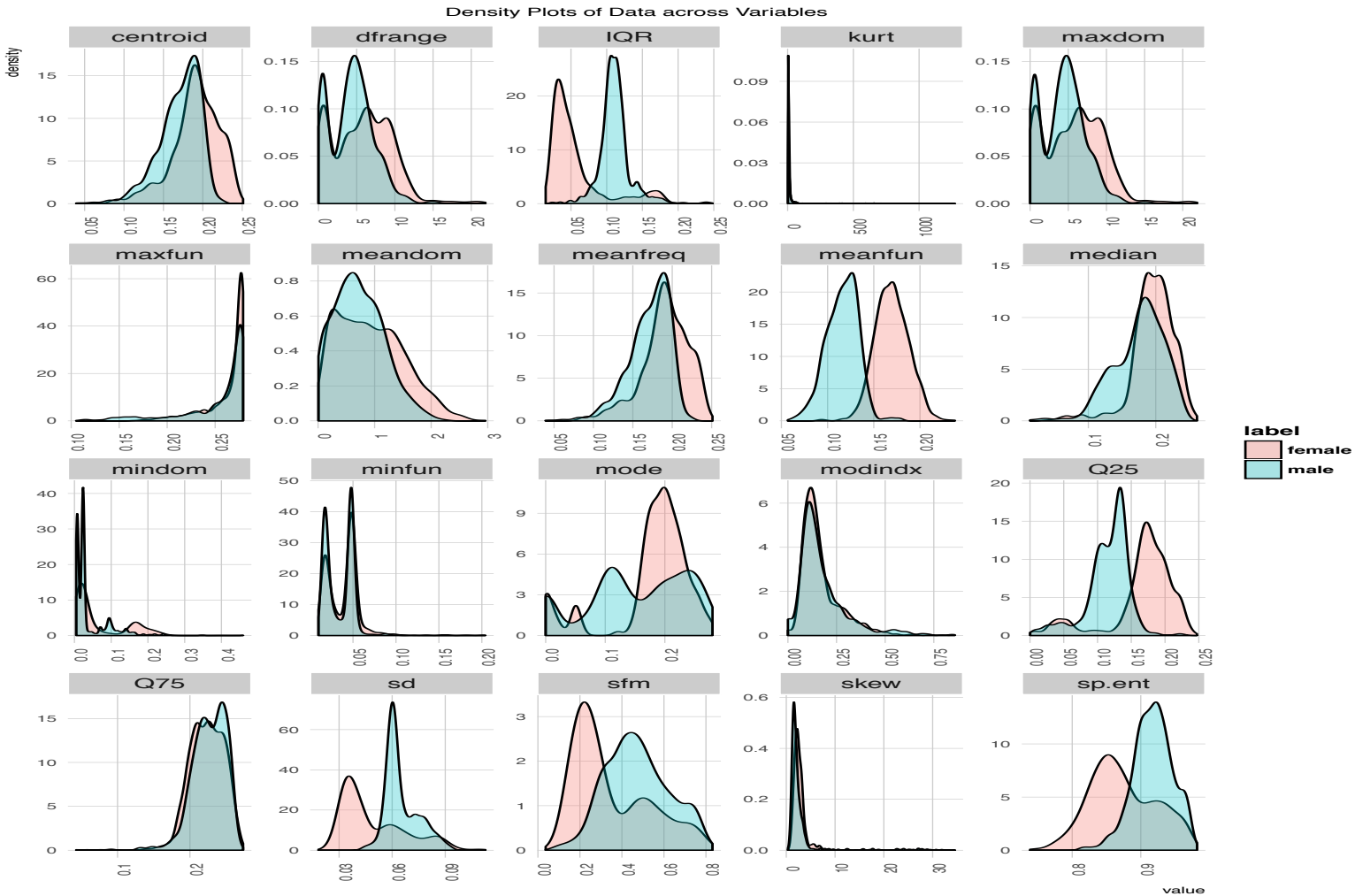


Overview of "Available" Data

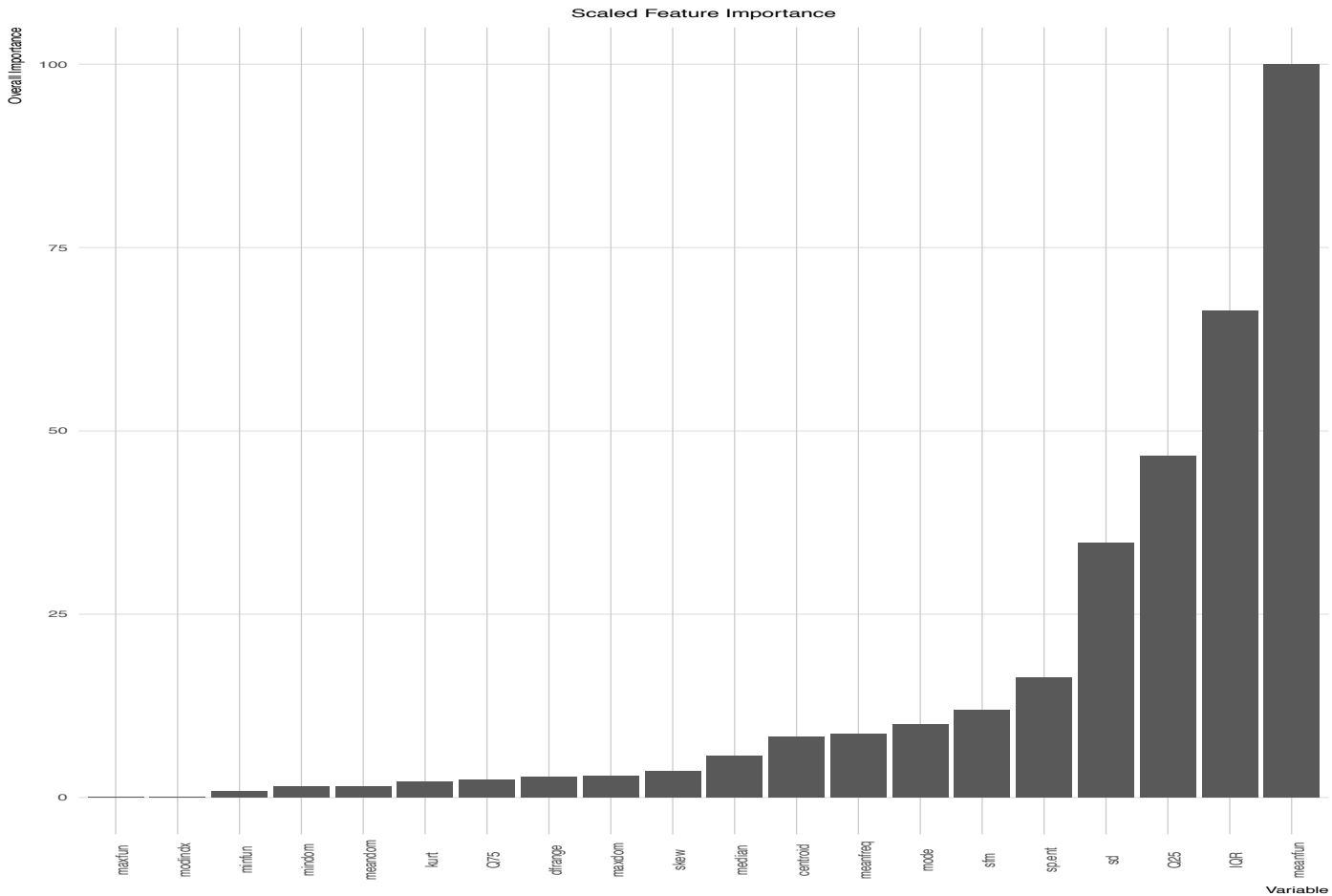


Overview of "Available" Data: Density

How does each numerical variable vary across the labels?

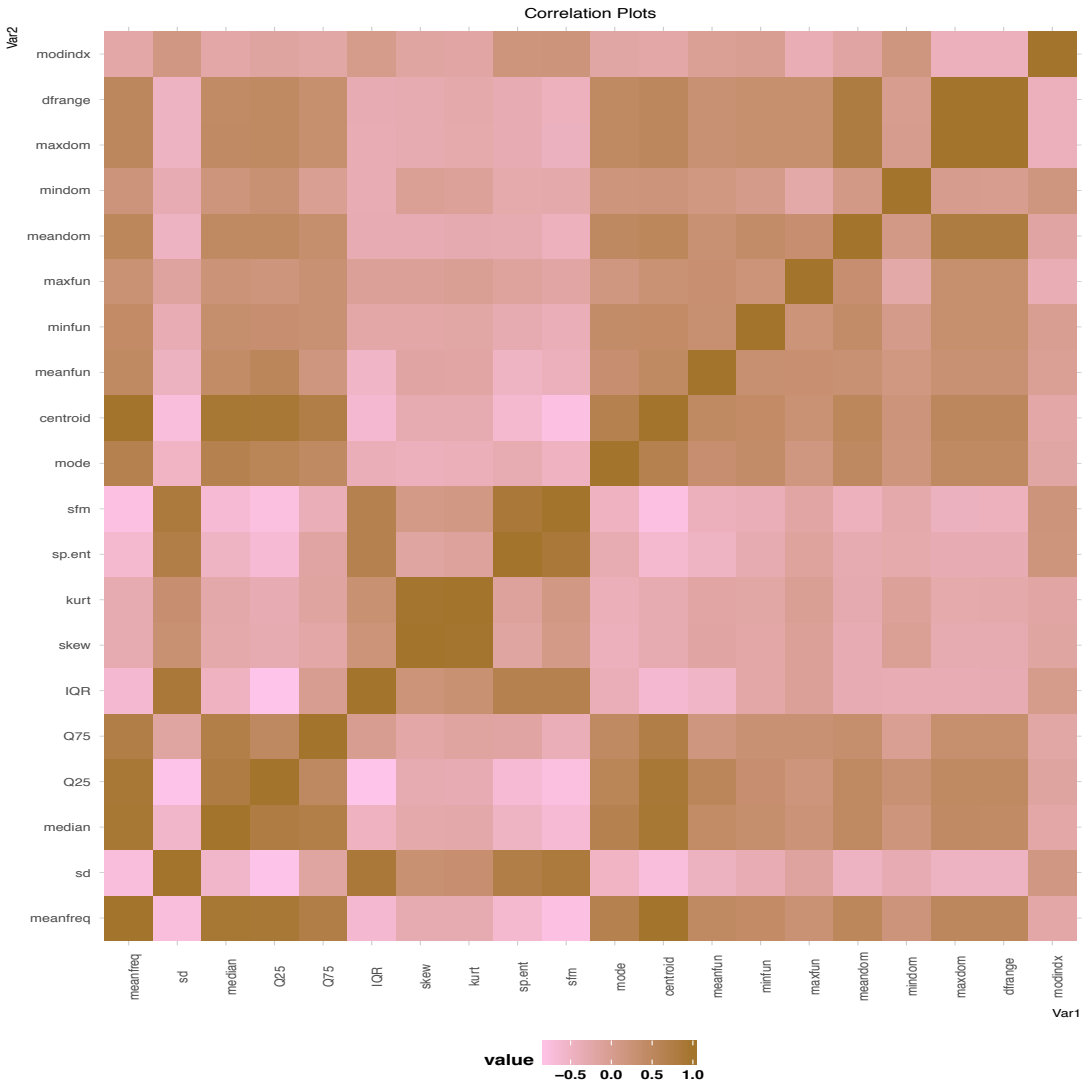


Overview of "Available" Data: Feature Importance

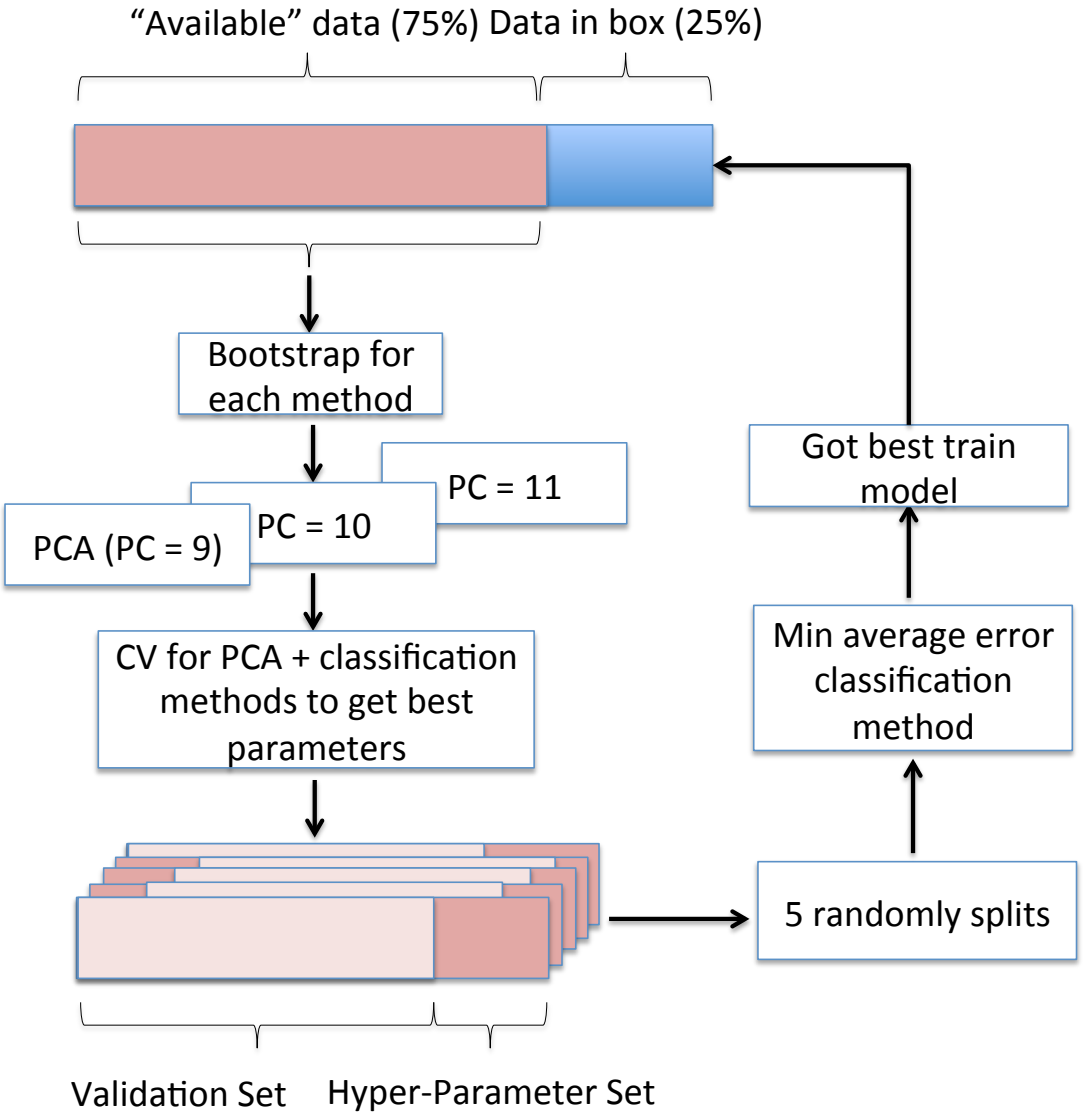


Overview of "Available" Data: PCA

Overview of "Available" Data: Correlated Variables



Flowchart and Classification Method



- Logistic Regression
- KNN
- Decision Tree
- Random Forest
- Boosting
- SVM

Results: Minimum Error for Each Method with Corresponding Method

Method	Minimum train error	Best principal components
Logistic Regression	0.0259	11
KNN	0.0057	11
Decision Tree	0.0603	10
Random Forest	0.0077	11
Boosting	0.1044	10
SVM	0.0131	9

Confusion Matrix for Training Set

LG	female	male
female	1429	48
male	29	1464

KNN	female	male
female	1492	6
male	11	1461

DT	female	male
female	1352	77
male	102	1439

RF	female	male
female	1509	12
male	11	1438

Boosting	female	male
female	1379	142
male	142	1307

SVM	female	male
female	1418	35
male	21	1496

Confusion Matrix for Testing Set

KNN	female	male
female	392	2
male	6	392
test error	0.01010101	

- KNN with K = 1 has the lowest average error rate on Hyper-Parameter set

RF	female	male
female	385	6
male	13	388
test error	0.0239899	

- RF has the second lowest average error on Hyper-Parameter set

Conclusion

- Each model had a satisfied error rate which is about or below 0.10

Conclusion

- Each model had a satisfied error rate which is about or below 0.10
- KNN and random forest generated the most optimal results for our training set, which gave testing error less than 0.01 (0.0057 for KNN and 0.0077 for random forest) and showed balanced predictions in the confusion matrix

Conclusion

- Each model had a satisfied error rate which is about or below 0.10
- KNN and random forest generated the most optimal results for our training set, which gave testing error less than 0.01 (0.0057 for KNN and 0.0077 for random forest) and showed balanced predictions in the confusion matrix
- As we tested the two models with testing set, KNN and random forest gave testing error 0.0101 and 0.0240

Conclusion

- Each model had a satisfied error rate which is about or below 0.10
- KNN and random forest generated the most optimal results for our training set, which gave testing error less than 0.01 (0.0057 for KNN and 0.0077 for random forest) and showed balanced predictions in the confusion matrix
- As we tested the two models with testing set, KNN and random forest gave testing error 0.0101 and 0.0240
- Based on the comparison, we would like to suggest to use KNN as the classification models in the similar voice-based gender identification issues

Limitation

- Nested CV for some models are time-consuming, thus parameters are assumed

Limitation

- Nested CV for some models are time-consuming, thus parameters are assumed
- Failed to implement LDA and QDA since dataset is not Gaussian-distributed

Limitation

- Nested CV for some models are time-consuming, thus parameters are assumed
- Failed to implement LDA and QDA since dataset is not Gaussian-distributed
- Neutral-gender like voice could be hard to identify

Reference

[1] https://www.kaggle.com/primaryobjects/voicegender

[2] James Gareth, Daniela Witten, and Trevor Hastie. "An Introduction to Statistical Learning:
With Applications in R." (2014)

[3] https://www.medcalc.org/manual/logistic_regression.php

[4] home.etf.rs/~vm/os/dmsw/Random%20Forest.pptx

Gender Recognition by Voice

DS 502 Final Project

Group Members:

Fangzheng Sun

Huimin Ren

Shanhao Li

Sahil Shahani

Yun Yue

Worcester Polytechnic Institute

April, 2016