CS 584 Final - Fall 2017

Jake Whalen

November 21, 2017

# Executive Summary

Summary of purpose/results

# Motivation

# Data

## Overview

There are +3,000 labelled records. The data contains 20 Features such as mean frequency, skewness, spectral entropy and average of fundamental frequency measured across acoustic signal. The goal is to predict whether a voice is male or female. A follow up goal is to determine what characteristics help distinguish between genders in the dataset.

The data was downloaded from <https://www.kaggle.com/primaryobjects/voicegender>.

# Approach

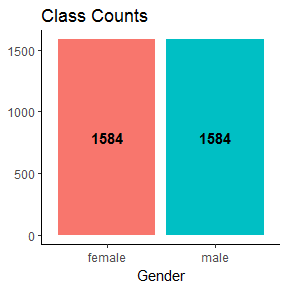
EDA

Machine Learning

Evaluation

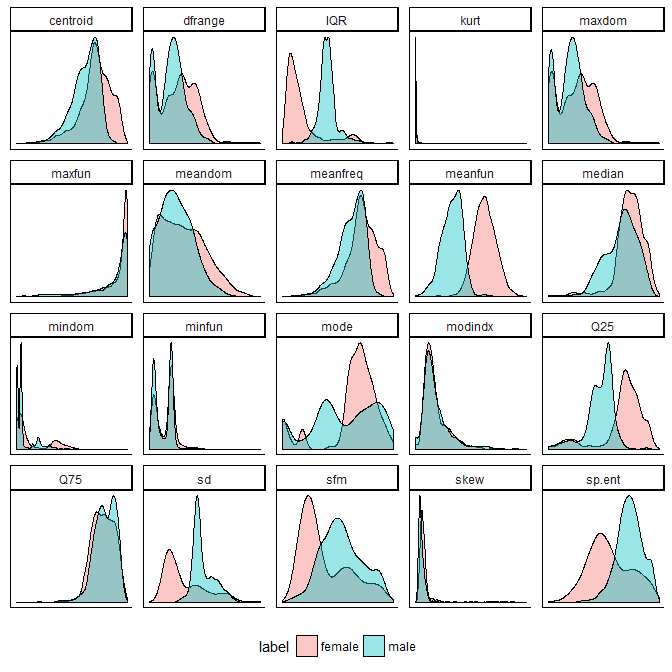
# Exploratory Data Analysis

#### Classes

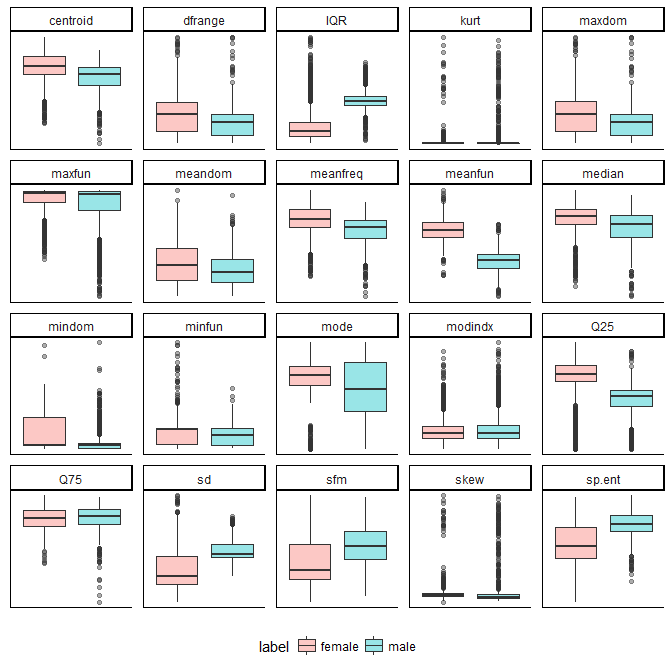


#### Visualize

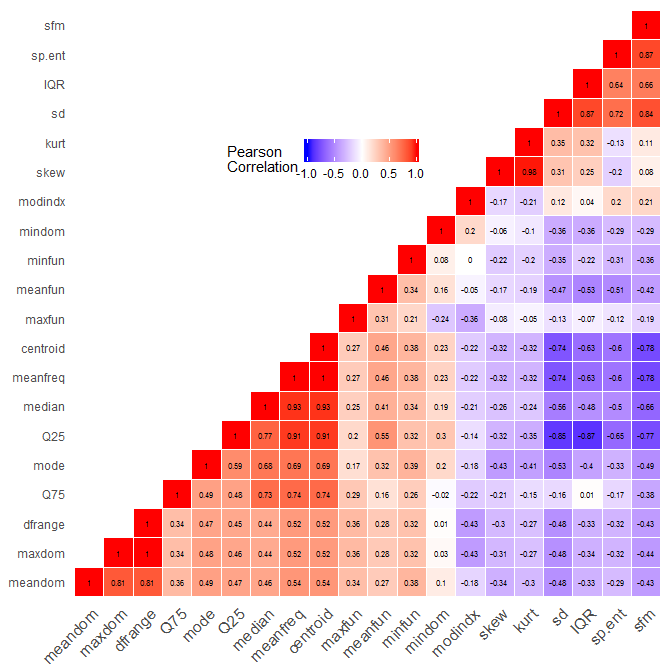
###### Distributions



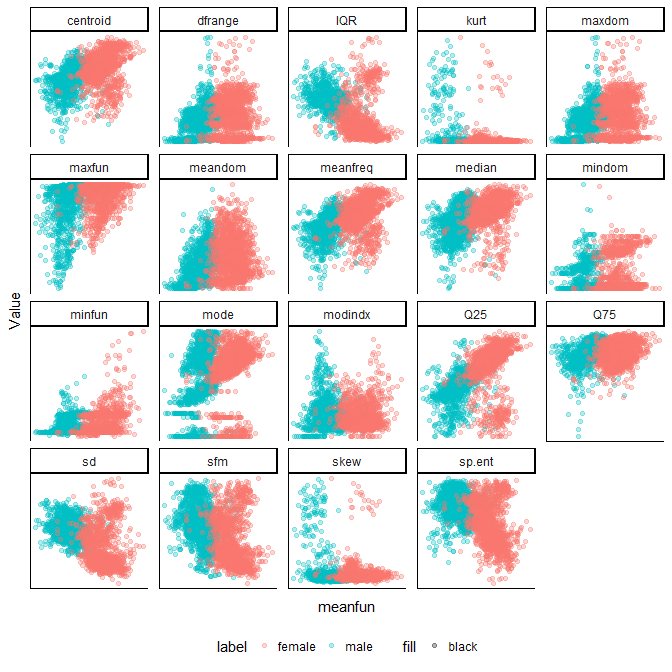
###### Boxplot



###### Heatmap



###### Scatter Plot



Also discuss 3D

#### Statistical Testing

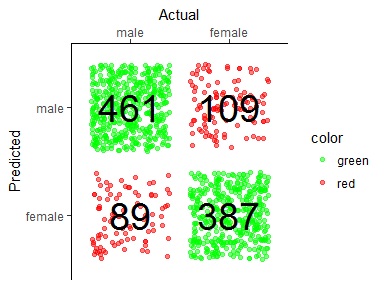
# Classification

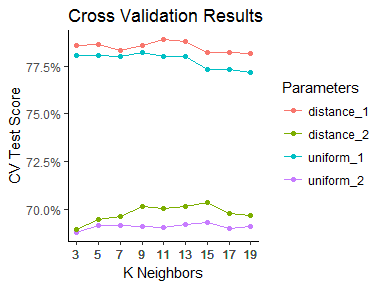
#### Machine Learning

###### Python

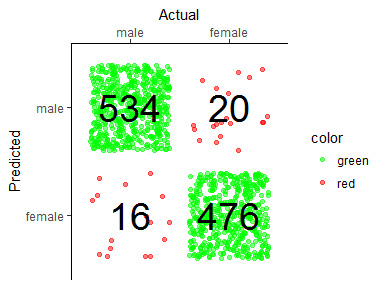
###### Scikit Learn

#### K-Nearest Neighbors

Confusion Matrix 

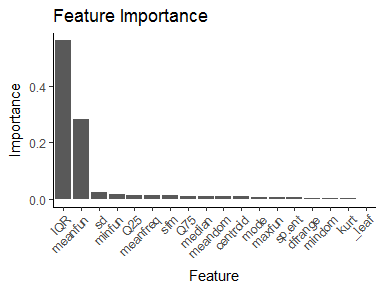
Cross Validation Results 

#### Decision Tree

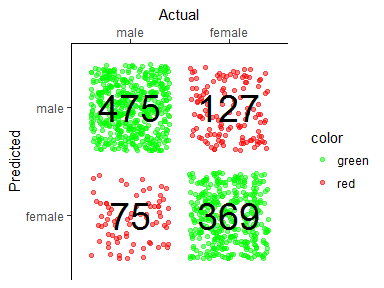
Confusion Matrix 

Feature Importance

## Warning: Removed 1 rows containing missing values (position\_stack).



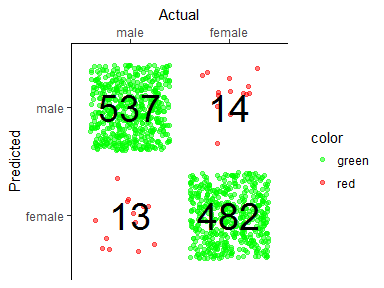
#### Support Vector Machine

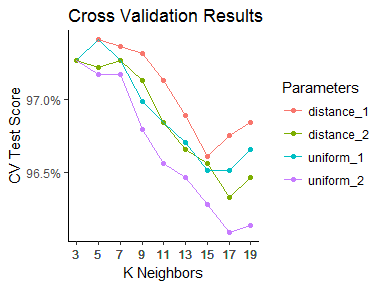
Confusion Matrix 

#### Logistic Regression

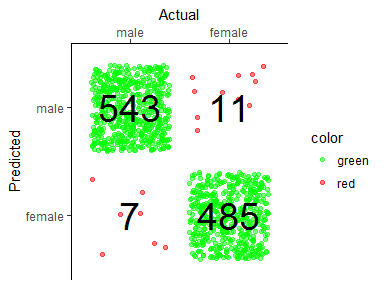
Confusion Matrix

#### K-Nearest Neighbors (PCA)

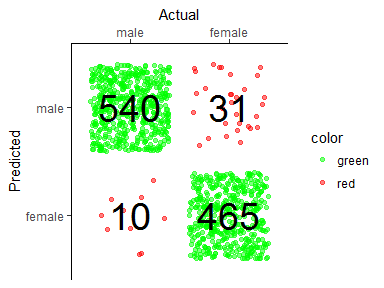
Confusion Matrix 

Cross Validation Results 

#### Random Forest

Confusion Matrix 

#### Support Vector Machine (PCA)

Confusion Matrix 

# Conclusions