**Identify gender of a voice**

Lei Shan

Capstone I Proposal

**Motivations:**

The study of human voice recognition can be traced back to 1950s when the first baby talk system “Andrey” was invented from Bell Laboratory. However, Andrey only recognizes digits from single voice. Later on, IBM produced “Shoebox” with stronger ability to understand 16 English words. After several decades of development, voice recognition technology acquired tremendous improvement. The products, such as apple Siri, google home and Amazon Echo, can even interact with human and adapt to people’s lifestyle. This is mostly due to the state-of-the-art machine learning algorithm, as programming *per se* is not adequate to analyze the complexity of human voice. In this project, I would like to identify the gender of human voice by using logistic regression based on its binary nature.

**Dataset:**

The dataset consists of 3168 recorded voice samples, collected from male and female speakers. There are around 20 features of voices, such as mean frequency, peak frequency and interquantile range.

**MVPs:**

a. EDA for dataset: there are a lot of features in the dataset. It is crucial to sort out independent features and remove outliers. Find the most influential features that determine the gender.

b. Utilize logistic regression to predict the gender of a voice. Obtain proper threshold to reduce variance and bias.

c. Optimize the model by going beyond classic logistic regression(decision tree, CART and etc).

**Features of dataset:**

* 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: male or female