**Gender Voice Determination (Male or Female) Based on Voice - Characteristics**

**Abstract:**

With human hearing ability most of the times we correctly guess the gender of a person. If we want computer/cellphone to do the same task is little challenging. In this activity we are attempting to develop/build a model which will guess the gender based once voice- characteristics.

In this project I have explored various classification algorithms and various data processing techniques to build a model with good classification result.

1. **DATA GATHERING:**

This database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech. The dataset consists of 3,168 recorded voice samples, collected from male and female speakers. The voice samples are pre-processed by acoustic analysis in R using the see wave and tune-R packages, with an analyzed frequency range of 0hz-280hz.

Data: <https://www.mldata.io/dataset-details/gender_voice/>

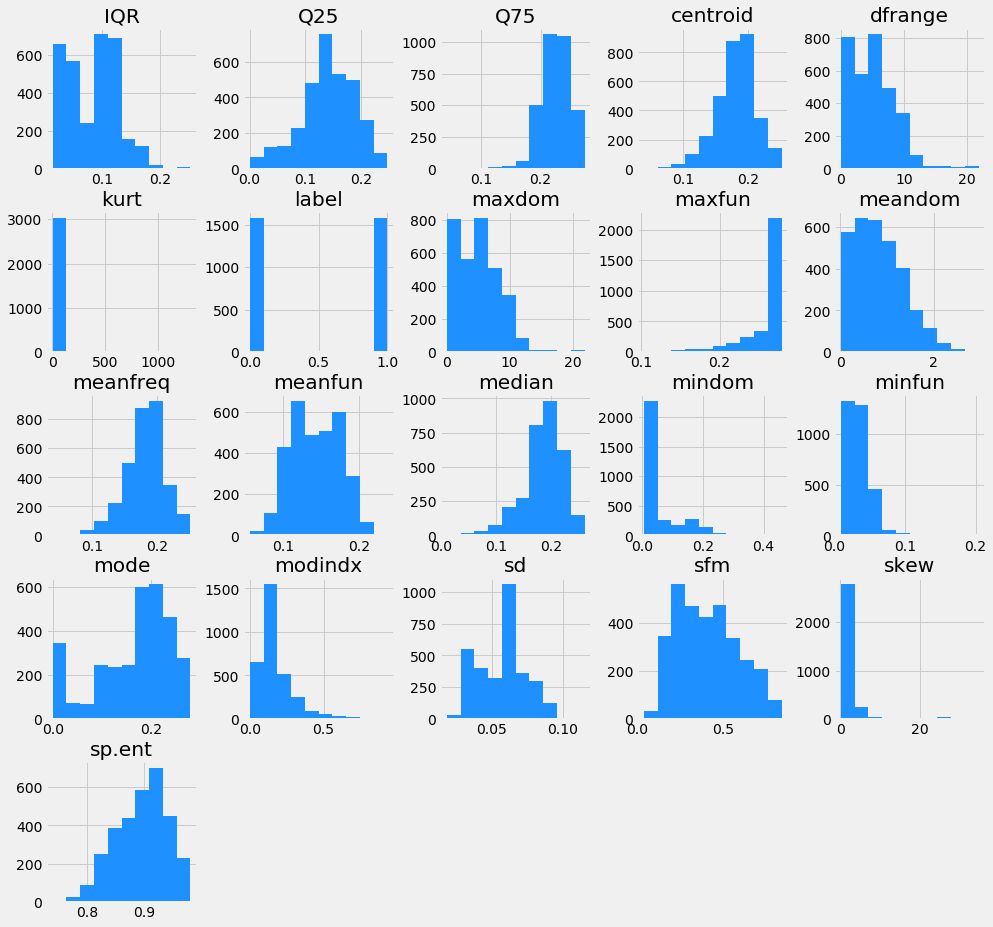
we have 3168 rows with 21 columns.

***Features:***

* 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

***Target:***

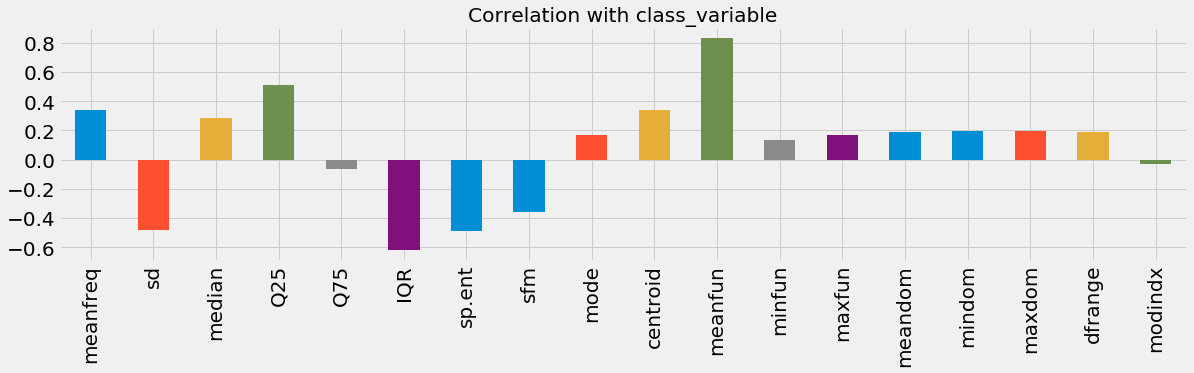
* *Label – male or female*

**

*From the histograms, we can say that:*

1. *Skew and Kurt have all zeroes in them and thus not contributing in our prediction. So, we can remove them from our dataset.*
2. *If we want remove more features or add then we need to see the learning curves and trail run result.*

*In the below graph we can see the relationship between predicting variables with target variable:*

**

1. **RESEARCHING THE MODEL THAT WILL BE BEST FOR THE TYPE OF DATA**

I have used following classification algorithms to get initial algorithm performance and insights of data by getting learning curves, classification report and ROC curves.

1. Logistic Regression Classifier

2. Decision Tree Classifier

3. Random Forest Classifier

4. Support Vector Machine Classifier

5. Gaussian NB Classifier

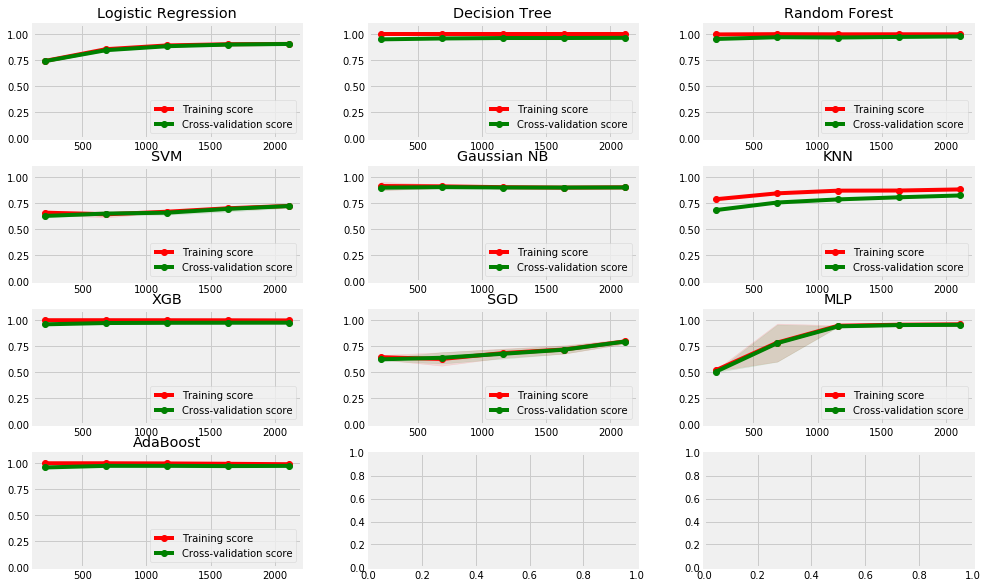
6. K -nearest neighbors Classifier

7. Gradient Boosting Classifier

8. Stochastic gradient descent

9. Multilayer perceptron

10. Adaboost



As per above graph, most of the estimators are biased but they have high classification accuracy. Decision Tree, Random Forest, XGB, MLP and Adaboost have more than 97% accuracy. We wanted to explore more and wanted to see if we can get 100% classification result. So, we tried to optimize the algorithms by using grid-search method. But there was no improvement in the classification result after using tuned models.

1. **MODEL PERFORMANCE ANALYSIS**

Firstly, we tested base and optimized models on validation dataset to get insights about the dataset and to get basic idea about used estimators. On testing dataset, XGB and Random forest outperformed all other algorithms. So, we finalized XGB and Random forest algorithms as our final models.

**Base run result:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 | ROC | Time |
| Logistic | 0.900631 | 0.966912 | 0.829653 | 0.893039 | 0.900631 | 0.005552 |
| Decision Tree | 0.966877 | 0.986842 | 0.946372 | 0.966184 | 0.966877 | 0.012965 |
| Random Forest | 0.981073 | 0.990354 | 0.971609 | 0.980892 | 0.981073 | 0.030437 |
| SVM | 0.675079 | 0.719368 | 0.574132 | 0.638596 | 0.675079 | 0.461295 |
| Gaussian NB | 0.90694 | 0.921569 | 0.88959 | 0.905297 | 0.90694 | 0.001572 |
| KNN | 0.79653 | 0.788344 | 0.810726 | 0.799378 | 0.79653 | 0.004966 |
| XGB | 0.976341 | 0.980892 | 0.971609 | 0.976228 | 0.976341 | 0.240344 |
| SGD | 0.742902 | 0.723837 | 0.785489 | 0.753404 | 0.742902 | 0.011616 |
| MLP | 0.973186 | 0.983871 | 0.962145 | 0.972887 | 0.973186 | 0.805457 |
| Adaboost | 0.981073 | 0.990354 | 0.971609 | 0.980892 | 0.981073 | 0.177738 |

After parameter tuning:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Model | Accuracy | Precision | Recall | F1 | ROC |
| 1 | Decision Tree | 0.957413 | 0.955224 | 0.963855 | 0.95952 | 0.957093 |
| 2 | Random Forest | 0.976341 | 0.970326 | 0.98494 | 0.977578 | 0.975914 |
| 3 | XGB | 0.977918 | 0.970414 | 0.987952 | 0.979104 | 0.97742 |
| 4 | MLP Classifier | 0.9653 | 0.96131 | 0.972892 | 0.967066 | 0.964923 |
| 5 | AdaBoost | 0.966877 | 0.966967 | 0.96988 | 0.968421 | 0.966728 |

Based on aboveresults we decided to go with Random Forest and Gradient Boost Classifier.

1. **MODEL TESTING AND FINAL MODEL SELECTION:**

We tested Gradient Boosting classifier and Random Forest model on testing dataset and we got very close result from both the models.

Random Forest: 0.97949 accuracy, with total 13 misclassification instances Gradient Boosting Classifier: 0.976340 accuracy, with total 15 misclassification instances

1. **CONCLUSION**

We explored various algorithms, their optimization processed and data pre-processing to came up with best classifying model. Based on our exploration, we have finalized two models as our models namely, XGB and Random Forest model.

**Scope for Developments:**

One can explore other classification algorithms and also deep learning to see if they can get 100% accuracy. We are going to stop here as we are getting high classification accuracy but one can explore other algorithms and optimization techniques to get 100% classification accuracy.