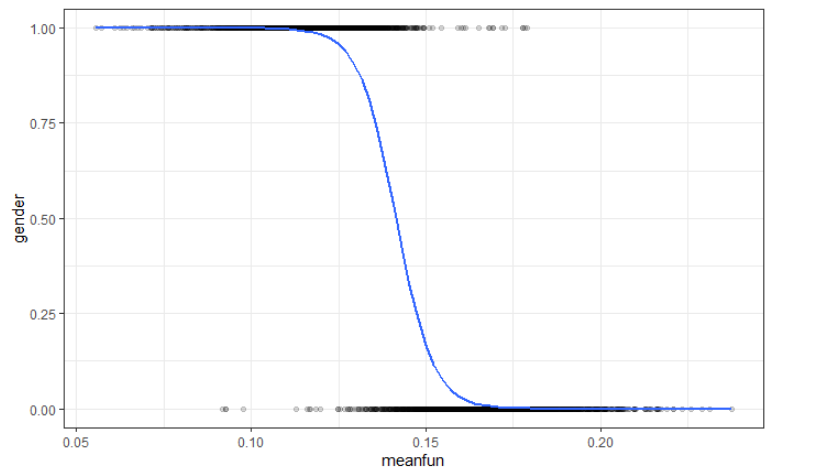
Diagram

Description automatically generated

**Fig. 1**

From the correlation plot (**Fig. 1**), it shows that there are potentially certain variables that have correlation with other variables.



**Fig. 2**

The above graphs from **Fig. 2** show a clear trend of logistic regression (sigmoid function) between meanfun and the gender. The meanfun variable corresponds to the average fundamental frequency measured across the acoustic signal, and from the plot, it can be observed that a voice with a meanfun above 0.14, the voice is likely to be a female voice. Apart from meanfun, there are other variables following the logistic regression trend and yet not show a clear trend as meanfun.

**Methods and Results.**

Identifying gender is the situation in which the response variable is male or female. Ideally, the appropriate models are logistic regression and random forest for classification. The categorical variable needs to be changed by creating a dummy variable and the column and its contents of gender were removed and replaced with the 1’s and 0’s. Then the dataset is split into train and test cases, 80% of the data was used for validating the model 20% was used for testing. As a result, the training data had 2534 rows, and the testing data had 634 rows.

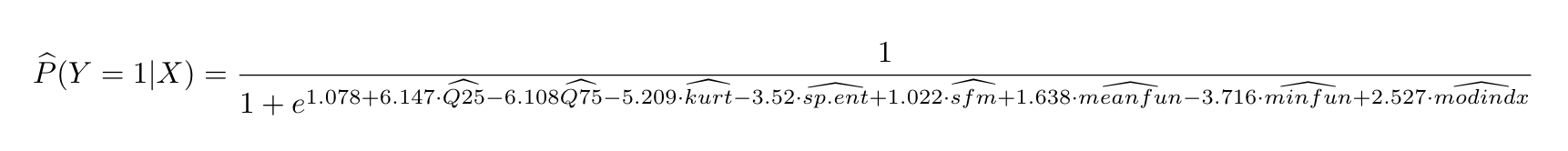
**Logistic Regression and Stepwise selection.**

The basic logistic regression analysis begins with logit transformation of the dependent variable through the utilization of maximum likelihood estimation. After applying the full logistic regression, the model was applied a backward stepwise selection to remove all the insignificant variables from the model, step () function used the AIC to pick out the best model with the optimized variables. As a result, our 21 predictors wered prune into 8 variables that are Q25, Q75, kurt, sp.ent, sfm, meanfun, minfun, modindx.

Table 3.1. Coefficients summary.

|  |  |
| --- | --- |
|  | Coefficients |
| Intercept | -10.78 |
| Q25 | -61.47 |
| Q75 | 61.08 |
| kurt | -0.052 |
| sp.ent | 35.2 |
| sfm | -10.22 |
| meanfun | -163.8 |
| minfun | 37.6 |
| modindx | -2.52 |

The probability of a gender male can be estimated with the following our final logistic regression model:



For the signs of the estimated coefficients, Q25, kurt, sfm, meanfun and modindx have negative coefficients, which can be interpreted as decreasing these predictor values will be associated with increasing the probability of being a male voice.

**Random Forest.**

The training and testing data are changed the label variables to factors and using random forest classification type. Using library randomForest for R code, the model is applied with the default number of tree ntree = 500, and number of variables tried at each split mtry = sqrt(p) = 4 where p is number of variables.

**Prediction results.**

Then the models are applied predicting on the testing dataset.

Table 4.2. Result of test data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Logistic regression full model** | | | **Stepwise selection** | | | **Random forest** | | |
| Gender | Correct | Incorrect | Gender | Correct | Incorrect | Gender | Correct | Incorrect |
| Male | 294 | 10 | Male | 295 | 9 | Male | 298 | 6 |
| Female | 322 | 8 | Female | 323 | 7 | Female | 324 | 6 |
| Total | 616 | 18 | Total | 618 | 16 | Total | 622 | 12 |
| Accuracy | 97.16% | | Accuracy | 97.47% | | Accuracy | 98.1% | |

The confusion matrix shows a predicting result on the Random Forest model with the highest accuracy at 98.1%. In terms of predictive performance, Random Forest is the best performance model. In terms of interpretability, Logistic regression with stepwise selection is the best choice for interpreting. Overall, the predicting performances of all three models are impressively good.