Assignment 3 - Part 2

Diagnosing schizophrenia from voice

Link to code:
https://github.com/StudiegruppeEM3/methods3_A3/blob/master/A3_P2_DiagnosingSchizophrenia_code.Rmd

Methods - models (Astrid)

We analyzed the data from 4 studies of speech patterns in Danish schizophrenic patients and controls. We chose to only use Danish speakers to keep the analysis simple. Schizophrenia may affect speech patterns differently in different languages. Furthermore when trying to predict schizophrenia in repeated measure models we decided to include study as a random intercept and not participant, this was due to the fact that including participant as a random intercept, explained an enormous amount of variance which basically made our acoustic features neglectable.
We made 10 models using mixed effect logistic regression. Each model predicted diagnosis from a single acoustic feature, with either name or study as a random intercept. We also included a full mode, with all 4 acoustic features. Then we compared these models using ANOVA where we looked for the model with the lowest AIC and BIC value. Furthermore, we made an importance plot of our 4 acoustic features (see figure 1). Both analyses indicated that pitch variability is our best acoustic feature.

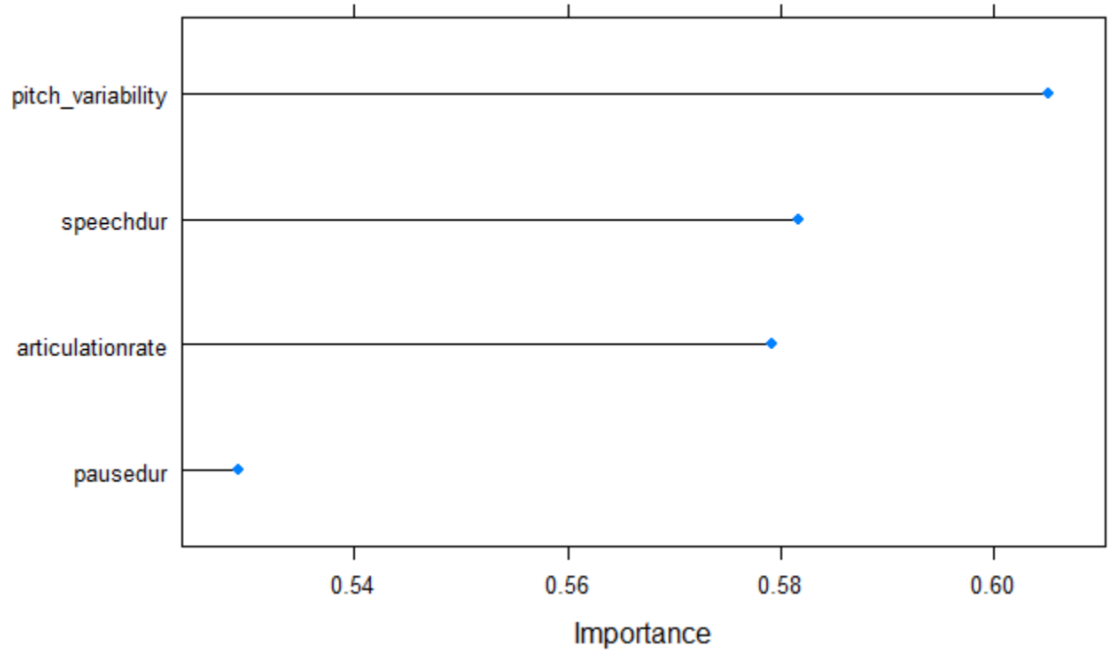

figure 1, shows the importance of our 4 acoustic features in predicting schizophrenia, results are 10-fold cross-validated, while the predictors have been scaled and centered. The metric used for importance was the area under the receiver operating characteristic curve (henceforth: ROC).

Methods - confusion matrix and ROC (Daniel)

We examined this model's (diagnosis ~ pitch-variability) confusion matrix and calculated the accuracy, sensitivity, specificity, positive and negative predictive value (henceforth: PPV and NPV). We also plotted the ROC (see figure 2). This was firstly done when training and testing on the same data set and afterwards done with a 10-fold cross validation.

Then, we calculated the accuracy, sensitivity, specificity, positive and negative predictive value, and plotted the ROC curve for the cross-validated model where every acoustic feature was included in the model. We did this for several different models to find the best possible model for predicting schizophrenia with our 4 features, we tried general logistic regression(GLM), elastic net logistic regression, repeated random forest, k-nearest neighbors(KNN), Support Vector Machine with a

Radial Basis Kernel Function(SVM), Linear Discriminant Analysis(LDA), Classification and Regression Trees (CART) and one ensemble model. Our ensemble model first consisted of GLM, KNN, SVM, CART and LDA however due to highly correlated models we excluded LDA and GLM because of correlation above 0.8, because highly correlated predictors can have detrimental effects on predictive power, see table 2. When stacking the ensemble model we used the method svmradial, because that was the best single best method (in respect to accuracy and area under the ROC) of the remaining 3 methods.

Correlation matrix for different classification methods

Model	IDA	CART	GLM	KNN	SVM
IDA	1	0.3	0.98	0.67	0.89
CART	0.3	1	0.35	0.003	0.46
GLM	0.98	0.35	1	0.69	0.92
KNN	0.67	0.003	0.7	1	0.5
SVM	0.89	0.46	0.92	0.51	1

tabel 2: Shows the correlation coefficients between the different models in the ensemble model.

Results (Jesper)

The best acoustic feature to predict schizophrenia diagnosis was pitch variability.

mean values of the model on pitch variability (Pernille)

Accuracy	0.58
Sensitivity	0.45
Specificity	0.69
PPV	0.57
NPV	0.58
Area under ROC	0.61
Kappa	0.15

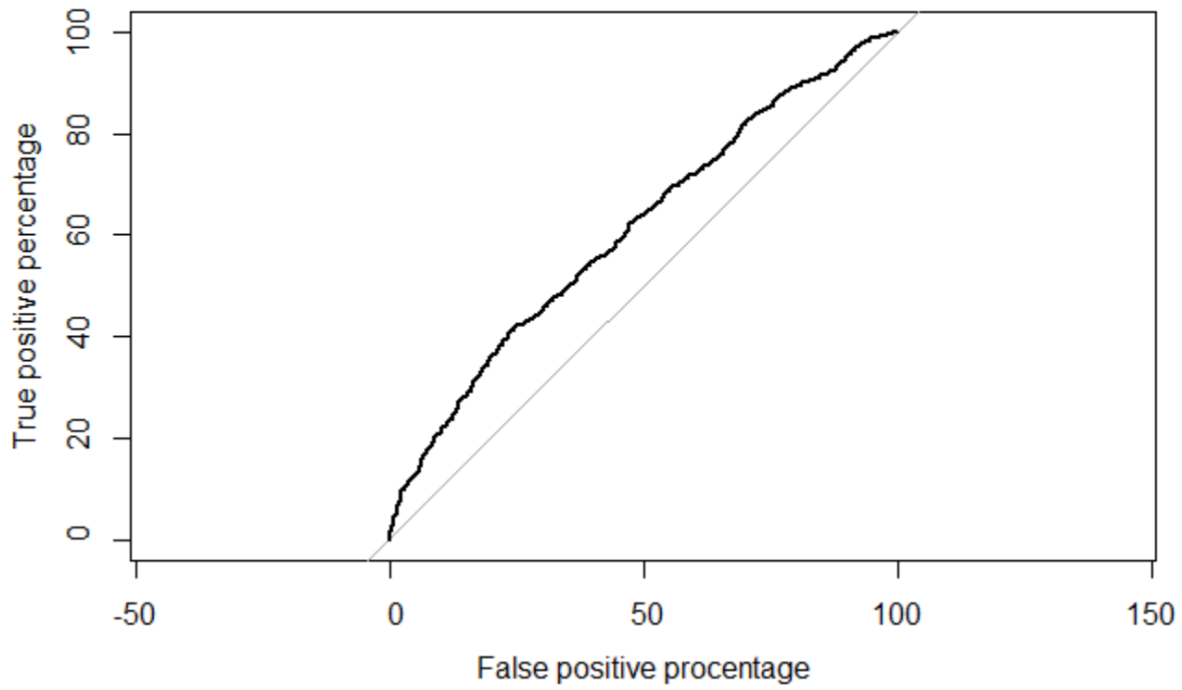

figure 2; ROC showing the relationship between 1-specificity in percent and sensitivity in percent for the model diagnosis ~ pitch variability

Mean values of the Cross-validated pitch-variability model. (Jesper)

Accuracy	0.54 (0.60-0.48)
Sensitivity	0.52 (0.61-0.44)
Specificity	0.56 (0.65-0.46)
PPV	0.52(0.59-0.48)
NPV	0.56 (0.63-0.49)
Area under ROC	0.60 (0.65-0.54)
Kappa	0.08 (0.21-(-0.05))

Mean values of several cross-validated models. (Jesper)

metric/ model	Accuracy	Sensitivity	Specificity	Kappa	Area under ROC
Glmer	0.58 (0.63-0.53)	0.49 (0.57-0.42)	0.67 (0.74-0.6)	0.16 (0.26-0.063)	0.61 (0.57-0.66)
elastic-net logistic regression	0.61 (0.65-0.56)	0.48 (0.54-0.42)	0.73 (0.79-0.67)	0.21 (0.31- 0.12)	0.61 (0.65-0.56)
Glm	0.60 (0.65-0.56)	0.49 (0.55-0.42)	0.72 (0.78-0.66)	0.21 (0.31-0.11)	0.60 (0.65-0.56)
CART	0.60 (0.62-0.58)	0.64 (0.71-0.58)	0.55 (0.59-0.5)	0.20 (0.24-0.15)	0.62 (0.65-0.60)
KNN	0.58 (0.61-0.56)	0.65 (0.66-0.63)	0.48 (0.52-0.45)	0.16 (0.21-0.11)	0.61 (0.63-0.58)

SVM	0.62(0.65-0.60)	0.68 (0.71-0.65)	0.54 (0.58-0.50)	0.24 (0.29-0.18)	0.65 (0.68-0.63)
LDA	0.61 (0.63-0.58)	0.71 (0.74-0.68)	0.50 (0.47-0.53)	0.21 (0.26-0.16)	0.63 (0.65-0.61)
Ensemble of KNN, SVM and CART	0.62 (0.64-0.60)	0.66 (0.69-0.63)	0.57 (0.62-0.5)	0.23 (0.27-0.19)	0.64 (0.66-0.61)

table 3. Showing different performance measures, as mean (upper-95%-confidence interval- lower-95%-confidence interval), from different models. The Glmer is with study as a random intercept

measure/ model	df	t-value	p-value
Glmer	9	4.67	0.0012
elastic-net logistic regression	9	4.60	0.0013
Glm	9	4.28	0.002
CART	9	10.64	0.000002
KNN	9	8.01	0.000022
SVM	9	13.8	0.00000023
LDA	9	14	0.0000002
Ensemble of KNN, SVM and CART	9	12	0.00000066

table 4 shows measures from a one-sided t-tests performed on areas under the ROCs for the given model on the values compared to a mean of 0.5. (see discussion to why we compare the values to 0.5) all p-values reported are uncorrected for multiple comparison.

Discussion

Discuss the "classification" process: which methods are you using? (Astrid)
For the first model we analyzed, we tested and trained the model on the same data. This is a problem because it leads to estimates that are unreliable if we use that model for predicting new data, that the model has not seen yet. We therefore continued to use a 10-fold cross-validation on every model made afterwards. As expected our estimates when cross-validating the pitch variability model, became significantly worse, but probably more accurate, if trying to predict new data.

Which confounds should you be aware of? (Daniel)
Some of the variables in our full model, including all acoustic features may be correlated. Normally, when we're trying to explain a phenomenon this would be a problem, since many models assume absence of multicollinearity. Since we're only trying to predict a diagnosis, this is only a problem if the prediction is worsened.
We checked to see if the model's predictors are correlated see table 5. They are not strongly correlated. If they had been, we would have checked how this affected the prediction of the models.

We also decided to scale and center all features used in every model, this can help with model convergence as well as processing time, one thing to be aware of when scaling and centering is that one has to scale the test data based on the scaling of the training data, else information about the test data will be leaked into the training data, ultimately skewing our estimates to be better than they actually are.

acoustic feature	pitch variability	articulation rate	pausedur	speech duration
pitch variability	1	-0.13	-0.05	0.06
articulation rate	-0.13	1	0.02	0.06
pausedur	-0.05	0.02	1	-0.14
speech duration	0.06	0.06	-0.14	1

table 5, showing the correlation matrix between the acoustic features.

What are the strengths and limitations of the analysis? (Jesper)

Our analysis shows that different models produce different results and can be used to solve problems, based on whether one thinks that sensitivity or specificity is more important. In our case we can see that the glm model produces models that have a high average specificity whereas the LDA- model produces models that have high average sensitivity. Our analysis only included Danish subjects which means that we do not know whether these estimates can be used to diagnose Chinese subjects, however from our representation of the data in assignment 3 part 1 we saw that the acoustic features did not have the same trend for Chinese individuals as compared to Danish individuals. Therefore, one can assume that even our best model predicting schizophrenia in the Danish subjects will do poorly in predicting schizophrenia in the Chinese subjects.
The biggest limitation of the analysis at hand is that the acoustic features are bad at predicting schizophrenia, values for the area under the ROC was at the highest 0.65 which could be categorized as being at least poor if not bad. This is because an area under the ROC of 0.5 corresponds to the model having no discriminatory ability at all. Therefore, when we compared our models to this value of 0.5, we evaluated if our models were better than randomly guessing.