Disclaimer: I understand that the final hand in should be more structured, and probably not as detailed as it is here for now (showing multiple plots for some models might be not necessary), but I am just wondering what is missing/incorrect in order to pass? Also, it took me a lot of time to grasp the new concepts, so the assignment is not fully finished yet.

Assignment 3

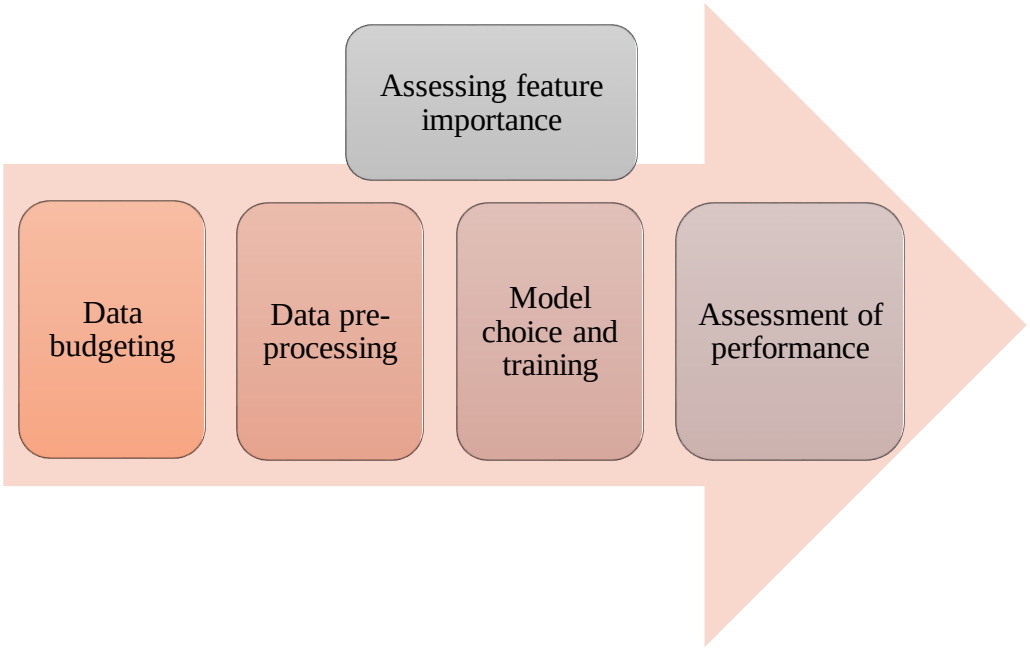
---

1. Describe your machine learning pipeline. Produce a diagram of it to guide the reader (e.g. see Rybner et al 2022 Vocal markers of autism: Assessing the generalizability of ML models), and describe the different parts: data budgeting, data preprocessing, model choice and training, assessment of performance.

2. Briefly justify and describe your use of simulated data, and results from the pipeline on them.

3. *Describe results from applying the ML pipeline to the empirical data and what we can learn from them.*

Overall machine learning pipeline



Data budgeting

The data is split into two parts: training (80% of the data) and testing (20 %) data. The split should preserve the structure of the data, meaning that participants from training data shouldn't appear in the testing data, there should be approximately equal amount of control and schizophrenic people in the training data in order for the algorithm to learn from both categories of participants equally. While the training set of data will be used for learning, the test set will be used to "verify" whether the algorithm could learn and infer the patterns.

Data pre-processing

The data (in our case, the simulated and empirical data) has to be scaled. The scaling of the data is performed only on the training dataset, as it would make sure that the information of heterogeneity of population would not affect the performance of the algorithm on the test set. Therefore, the algorithm will not "know" the full variability of the data, and this information would not help the algorithm to classify the data more correctly.

Model choice and training

In order to find the model that might work best with our data, I chose three models to begin with: model with fixed effects, varying intercepts, varying slopes. To assess the quality of each model (how well the model can classify a person belonging to the group of control or schizophrenia), the algorithm of logistic regression is used. Furthermore, I assessed how our priors affect the models' performance (sensitivity analysis of accuracy) to see whether the priors should be more conservative or looser.

Assessment of performance

How well the model performs the classification is assessed by looking at the accuracy estimate of classification, and seeing what kind of errors does the model make. It might be that the model classifies more "controls" as "schizophrenics", which might seem to be less significant error than as classifying "schizophrenics" as "controls".

Assessing feature importance

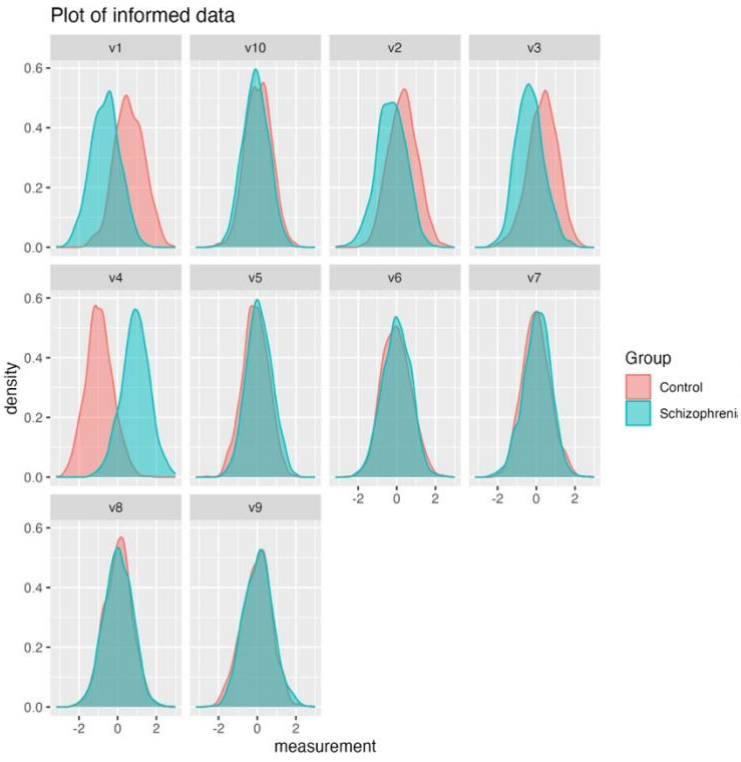
The feature importance is assessed by looking at the coefficients of each of the model, and by carrying out the analysis of global feature importance. The results indicate which features are used by the model a lot, and are most significant when classifying the sample into the groups of control and schizophrenia.

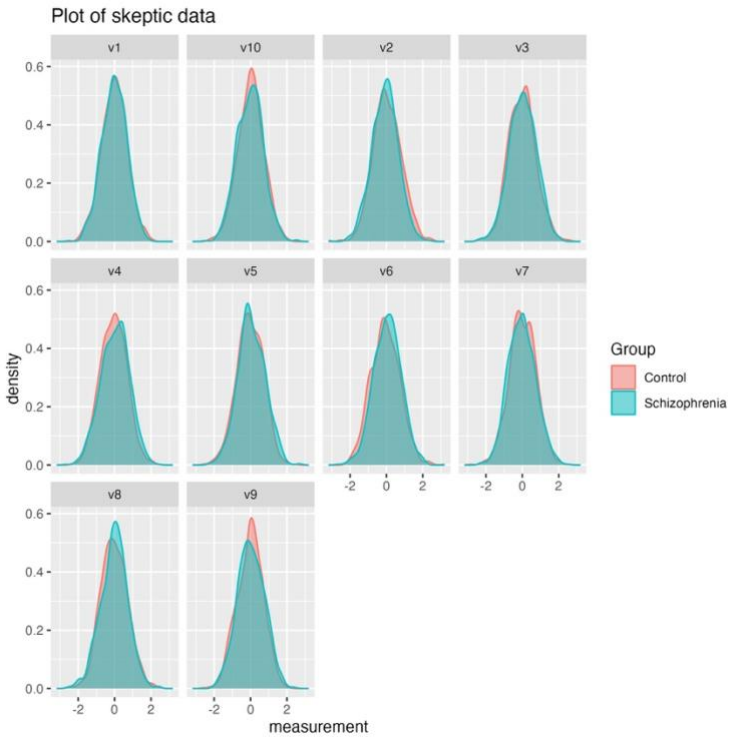
Question 2

Data simulation will allow to understand the problem of classification more, as I know what simulated data consists of and which information encoded in the data simulation is useful when separating people into control and schizophrenic. Later, the machine learning pipeline is used as "marker" to see whether the data patterns are inferred and whether, according to specific measures, participants can be classified correctly. Moreover, it will help to figure out which features impact the classification algorithm the most and also understand the results of empirical data.

Two datasets with 100 matched pairs of controls and schizophrenic patients were simulated. One dataset consists of 10 acoustic measures – noise variables ("skeptic" data), another one ("informed") includes 6 measures from the meta-analysis and 4 of random noise. The meaning of the measures from the meta-analysis is indicated below:

Acoustic measure	Proportion of spoken time	Pitch variability	Speech rate	Duration of pauses	Pitch mean	Number of pauses
Effect size	- 1.26	- 0.55	- 0.75	1.89	0.25	0.05





The plots of informed and skeptic data are shown above. The biggest difference can be seen in measures v1 – v4, as these have the biggest effect sizes, whereas measures v5 – v6 are closer to 0 and seem to overlap more. As effect sizes in v7 – v10 are just additional noise and are equal across both data sets, there is no clear difference in the plots as well.

Application of ML pipeline on simulated data sets

The data budgeting is performed on both informed and skeptic data sets. 80% of the data from each data set is used as a training data, the rest 20% as a test data. Moreover, it was made sure that the same participant would not appear in both training and testing data sets.

The measures in the training data sets of both skeptic and informed were scaled by using the mean and standard deviation of each feature (v1 – v10).

In order to check which features in classification problem matter the most, I set up three different models for informed and skeptic data sets separately: one consisting of fixed effects, one of varying intercepts and the last one of varying slopes:

Fixed effects: *Group ~ 1 + v2 + v1 + v3 + v4 + v5 + v6 + v7 + v8 + v9 + v10*

Varying intercepts: *Group ~ 1 + v2 + v1 + v3 + v4 + v5 + v6 + v7 + v8 + v9 + v10 + (1|ID)*

Varying slopes: *Group ~ 1 + v2 + v1 + v3 + v4 + v5 + v6 + v7 + v8 + v9 + v10 + (1 + v2 + v1 + v3 + v4 + v5 + v6 + v7 + v8 + v9 + v10|ID)*

Performance of the models with informed (scaled) training data set.

The top output is for model with fixed effects, the middle one – varying intercepts, the bottom one – varying slopes. It applies to all of the models below as well.

	Truth	
Prediction	Control	Schizophrenia
Control	761	38
Schizophrenia	39	762

	Truth	
Prediction	Control	Schizophrenia
Control	761	38
Schizophrenia	39	762

	Truth	
Prediction	Control	Schizophrenia
Control	800	0
Schizophrenia	0	800

Performance of the models with informed test data set.

	Truth	
Prediction	Control	Schizophrenia
Control	189	10
Schizophrenia	11	190

	Truth	
Prediction	Control	Schizophrenia
Control	188	10
Schizophrenia	12	190

	Truth	
Prediction	Control	Schizophrenia
Control	187	15
Schizophrenia	13	185

Performance of the models with skeptic (scaled) training data set.

	Truth	
Prediction	Control	Schizophrenia
Control	473	308
Schizophrenia	327	492

	Truth	
Prediction	Control	Schizophrenia
Control	468	311
Schizophrenia	332	489

	Truth	
Prediction	Control	Schizophrenia
Control	800	1
Schizophrenia	0	799

Performance of the models with skeptic test data set.

	Truth	
Prediction	Control	Schizophrenia
Control	110	92
Schizophrenia	90	108

	Truth	
Prediction	Control	Schizophrenia
Control	109	93
Schizophrenia	91	107

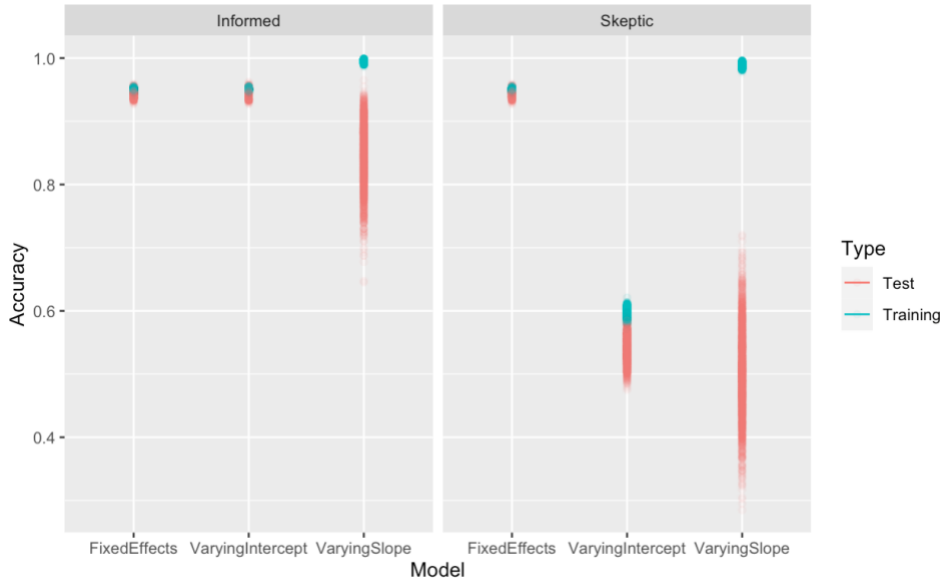
	Truth	
Prediction	Control	Schizophrenia
Control	112	87
Schizophrenia	88	113

In case of informed training data set: it seems like both – fixed effects and varying intercept – models on average perform similarly when classifying the patients into controls and schizophrenics (also have the same estimate of accuracy). The model with varying slope performs categorization perfectly (too perfectly), with estimate of accuracy equal to 1 (overfitting?).

In case of informed test data set: fixed effects model performed the best, as it classified the groups most accurately, while model with varying slopes performed poorer. Therefore, the models with the training data were way more accurate in classification of diagnosis compared to the test data – the data that the models weren't exposed to before.

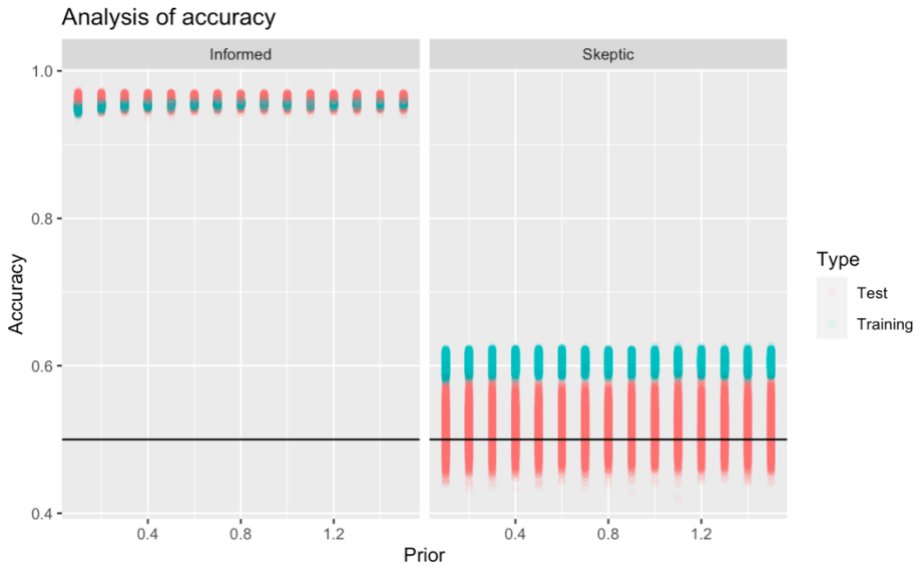
In case of skeptic training data set: the model with varying slopes performs the best, almost in the same way as with informed training data set.

In case of skeptic test data set: the model with varying slopes, again, performs the best, although all of the models classify people as control or schizophrenic poorly, just above the chance level (accuracy of ~ 0.5). In general, the pattern of model performance with skeptic data (the data with no effects) was pretty much similar to the models with informed data set.

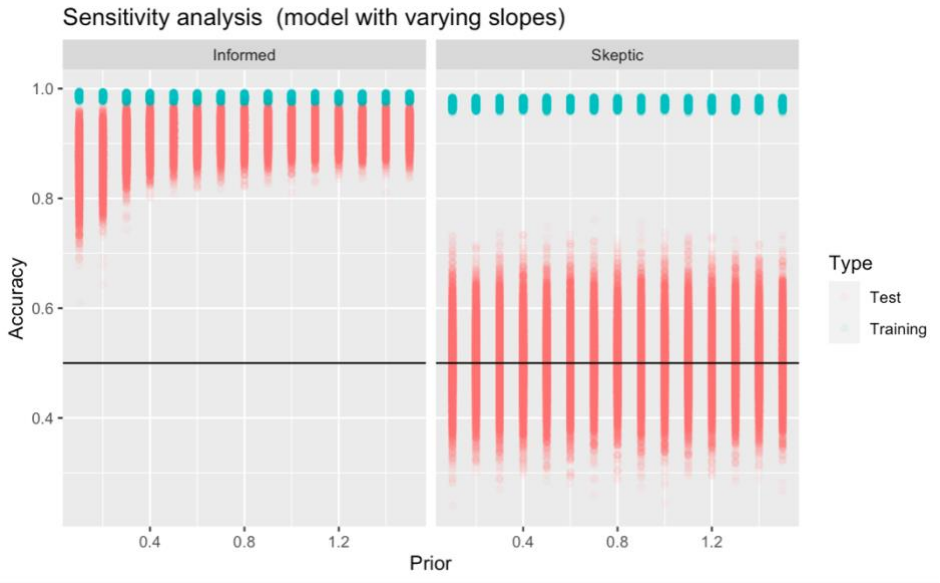


The figure above summarizes the results of accuracy when fitting different models. The performance of the fixed effect model with skeptic and informed data set seems to be similar. Not sure how to explain that?

Sensitivity analysis of accuracy (assessing the impact of priors)



The figure above indicates how priors impact the accuracy of the model when it classifies participants into the groups of diagnosis. Here, the performance of fixed effects model is captured. It seems that the prior does not really affect the accuracy of classification, as the uncertainty for both informed and skeptic data does not change across all of the values of the prior.



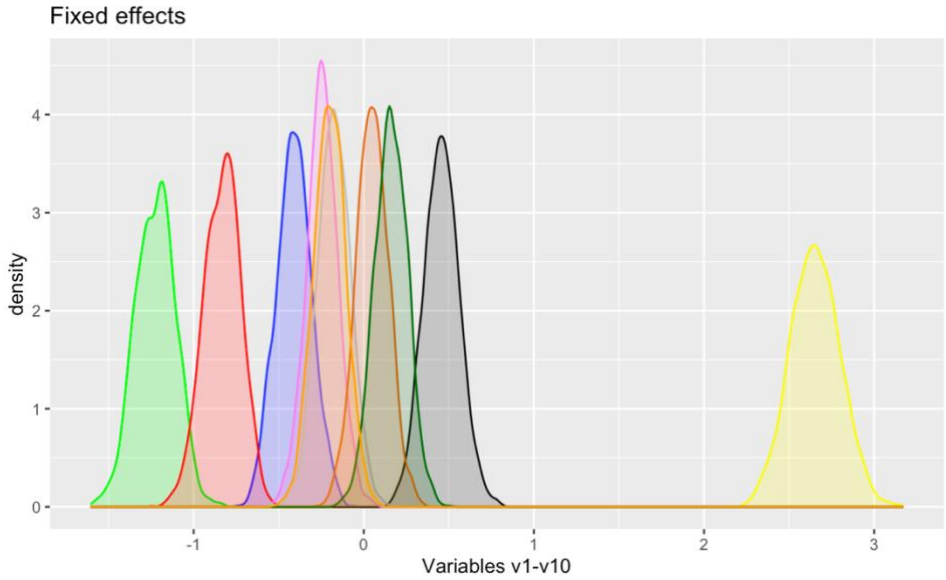
However, the results are different when it comes to setting priors for the model with varying slopes (above). Here, less conservative priors reduce the uncertainty (in the test set) of classification (with informed data set), nevertheless, do not affect the uncertainty with skeptic data set.

I did not run the sensitivity analysis on other models, as it takes ~3 hours to finish for one model. And not sure if it is needed in this case.

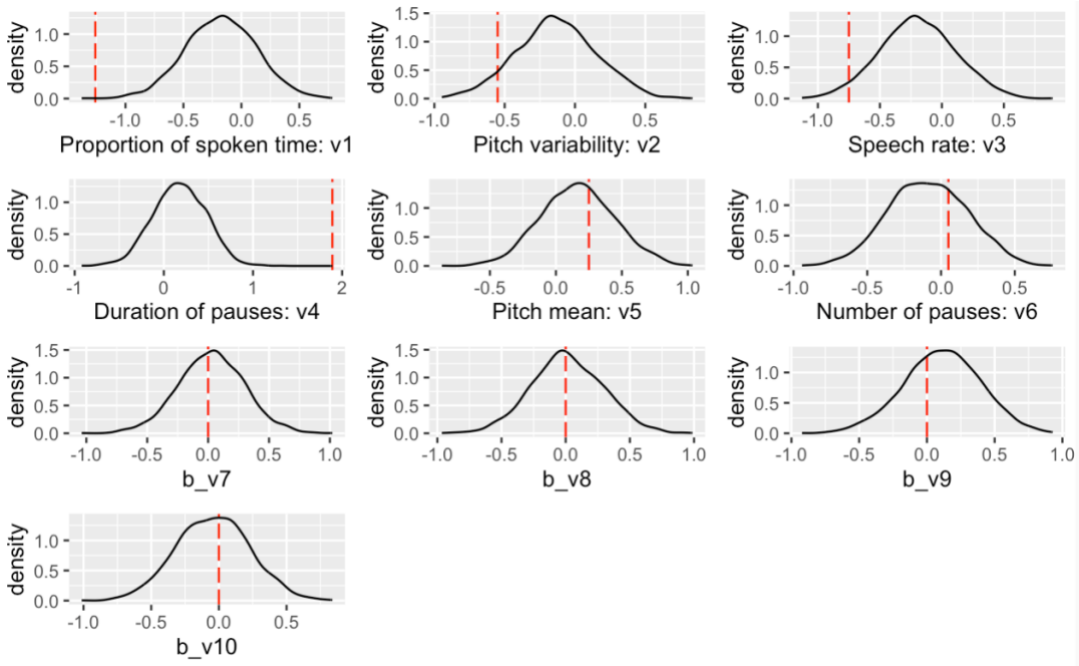
Feature importance

I have assessed the feature importance in all of the models with informed data set. Below, the results of each model are summarized.

Fixed effects model:

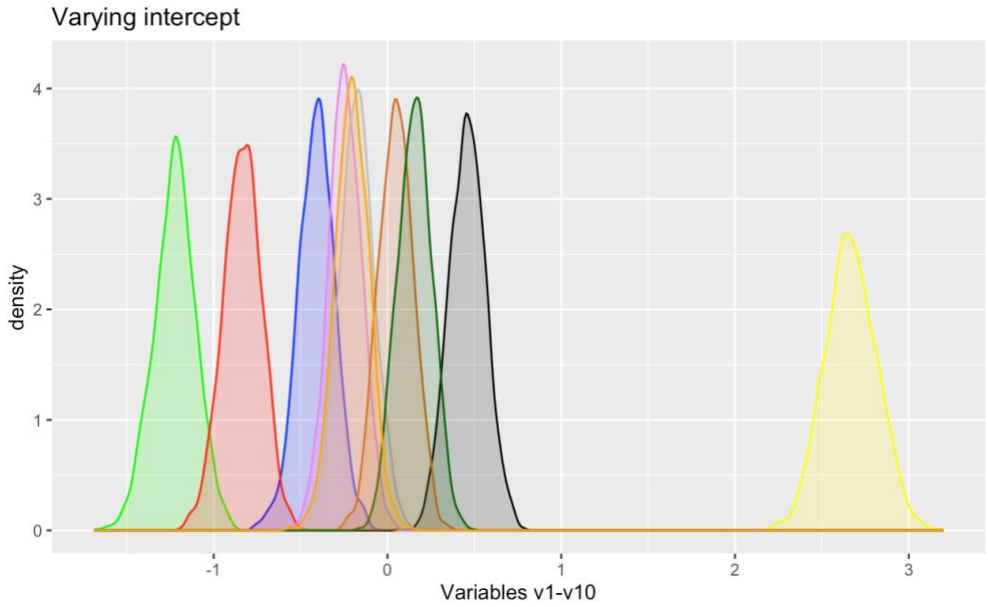


By looking at the posterior distributions of each of the variable, the duration of pauses (yellow), the proportion of spoken time (green), and the speech rate (red) are used a lot by the model.

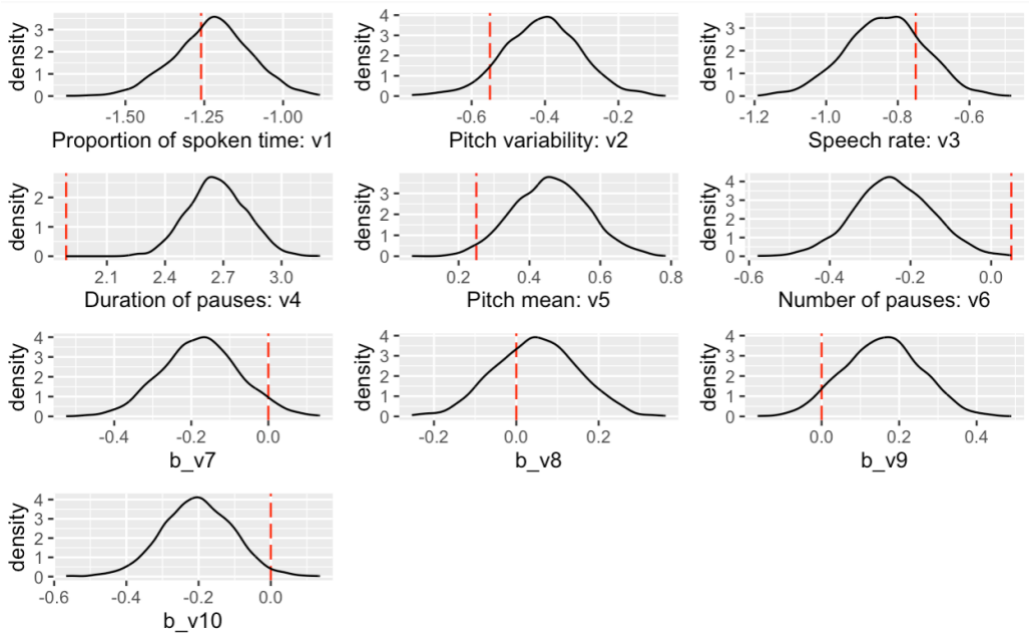


The density plots above illustrate the posterior distributions and whether they capture the true effect. How to explain why the v1/v4 – proportion of spoken time/duration of pauses – are used by the model a lot, but here it seems like it does not include the true mean?

Varying intercepts model:

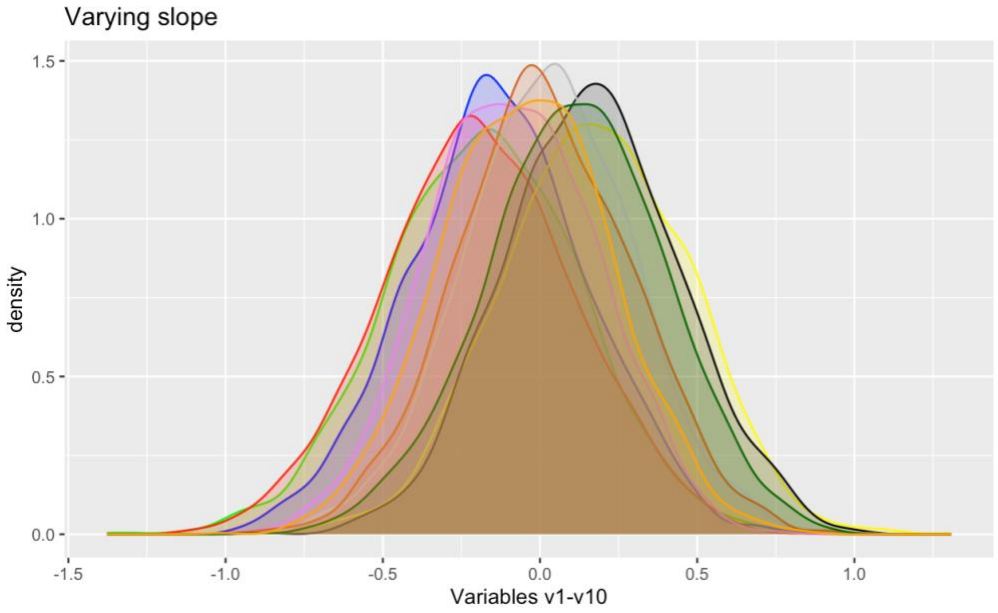


The output seems to be very similar compared to the model of fixed effects. The importance of the features is the same.

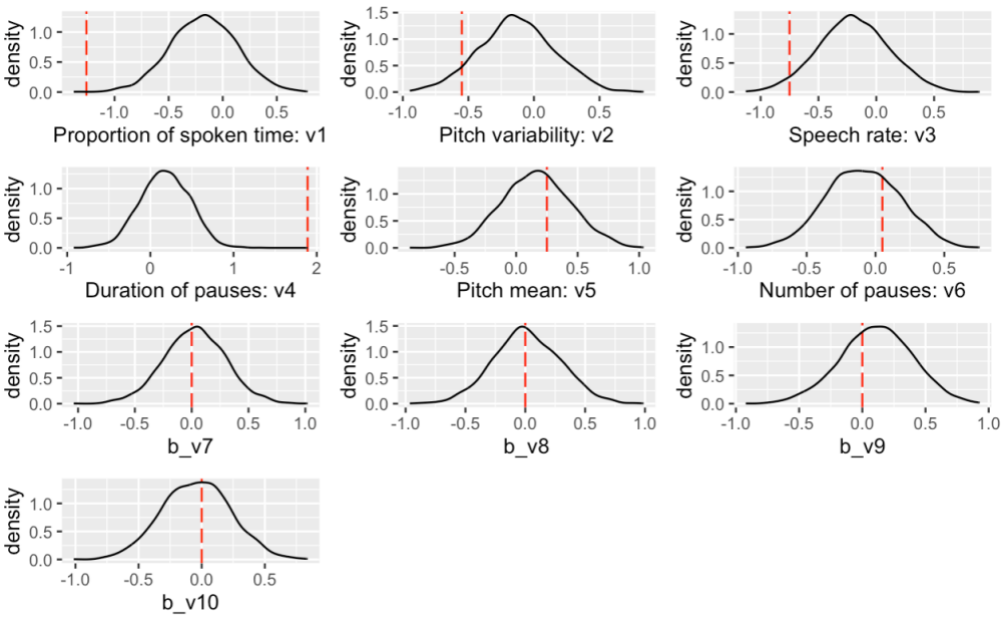


The density plots are also pretty similar to the ones shown above, however, in this case, the model with varying intercepts captures the true effect of v1 and v3 more accurately.

Varying slopes model:

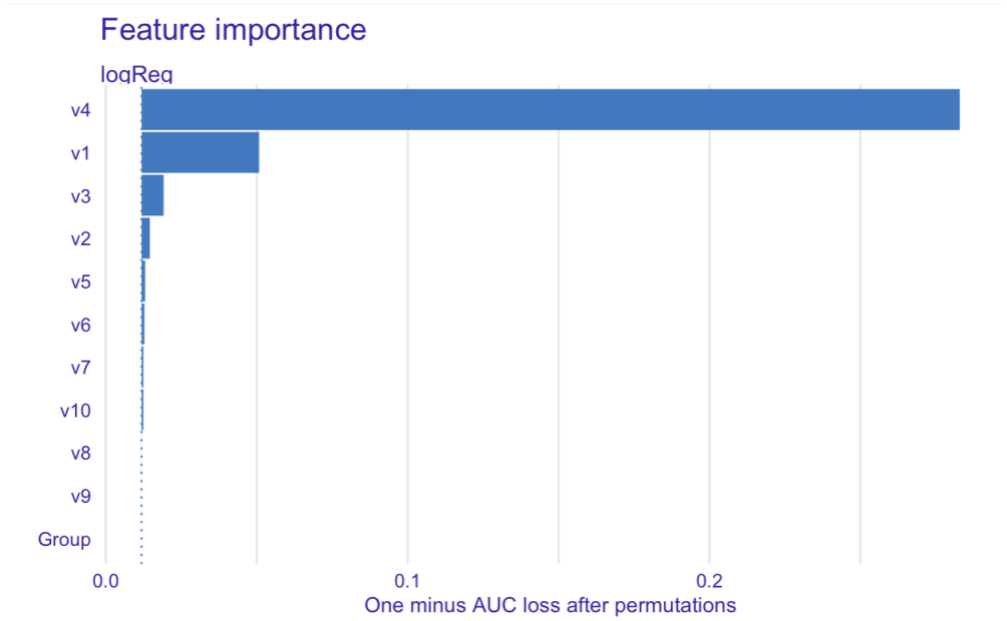


The distributions are overlapping a lot, the duration of pauses (v4 – yellow) becomes the important feature in this model (see the distribution of v4 below).



Here, I included the analysis of feature importance for the models separately, however, should I leave it for the final hand in? or the global feature importance check is enough to perform?

Global feature importance:



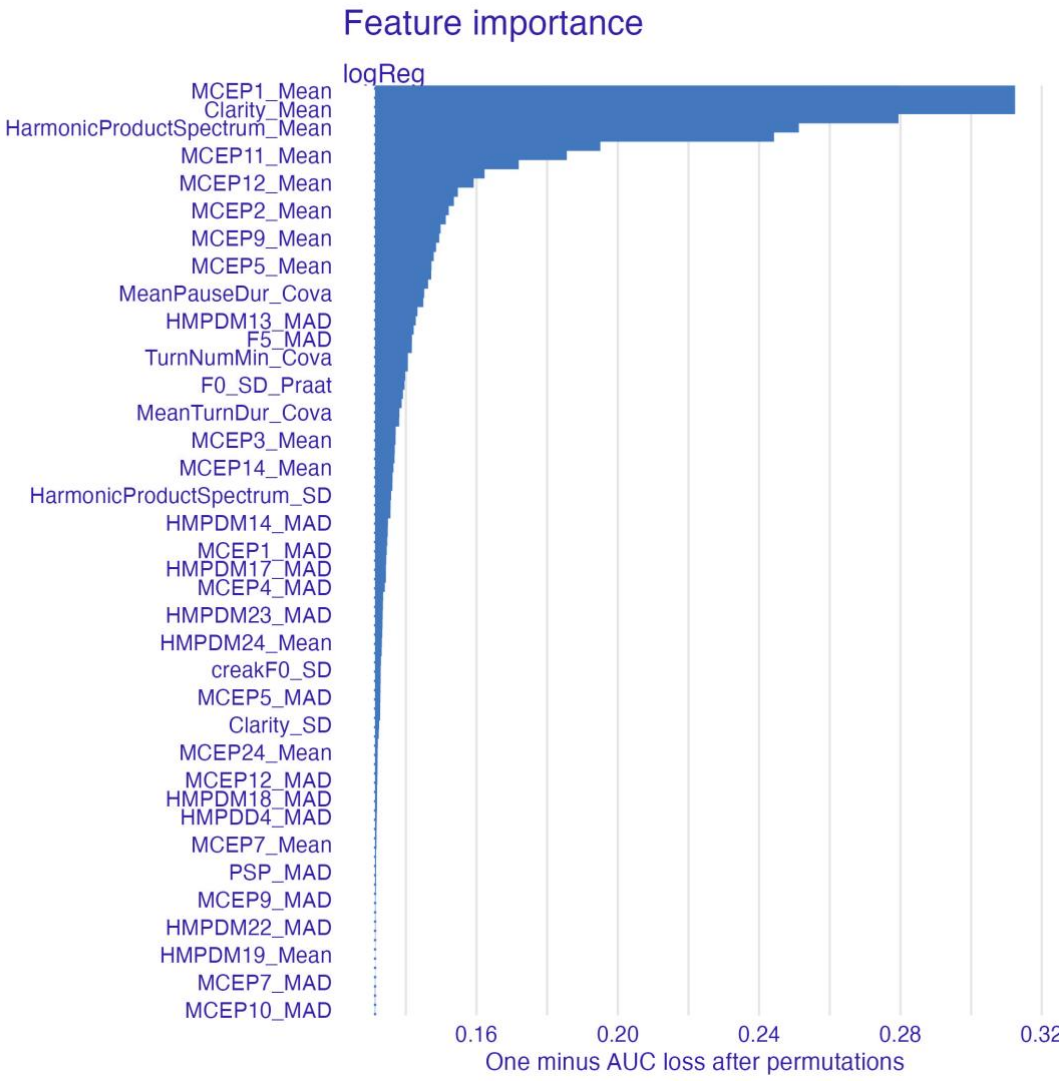
By using logistic regression algorithm so assess the feature importance, the results indicate that again, v4 is the most significant feature that is used, followed by the importance of v1 and v3. The general pattern of importance is similar to the one with models' coefficients described above.

The importance of the features that the model is estimating is plotted above. Therefore, as the duration of pauses increases (v4), the probability of being diagnosed as schizophrenic increases as well. Moreover, as the proportion of spoken time decreases (v1), the probability of being diagnosed as schizophrenic decreases too. The measure of speech rate (v3) follows the same pattern as the proportion of spoken time; however, it is not as significant measure as the latter one is. From this plot, the "noise" measures v7-v10 might seem to show the importance to the model, but as the effect of these measures is 0, and also, plots are generated on a smaller scale

compared to the measures v1-v6, they are just pure noise which predict the diagnosis at the chance level.

Question 3

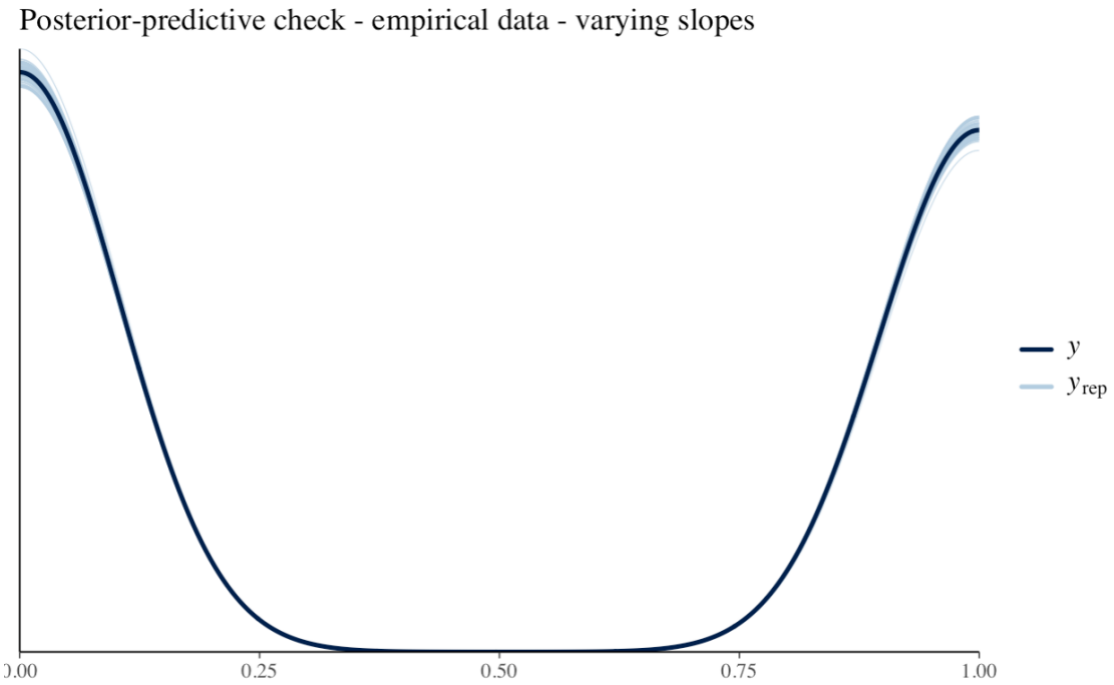
Now, the machine learning pipeline will be applied to the empirical data. At the very beginning the empirical data (as it is) was used to see which features of the data matter to the model the most. However, R could not handle the amount of the variables that the data consists of, therefore certain features had to be cut down. In this case, I have removed non-acoustic features, such as IDs, gender, language, etc… Moreover, the features, that are highly correlated (correlation > 0.7) were also removed from the dataset.



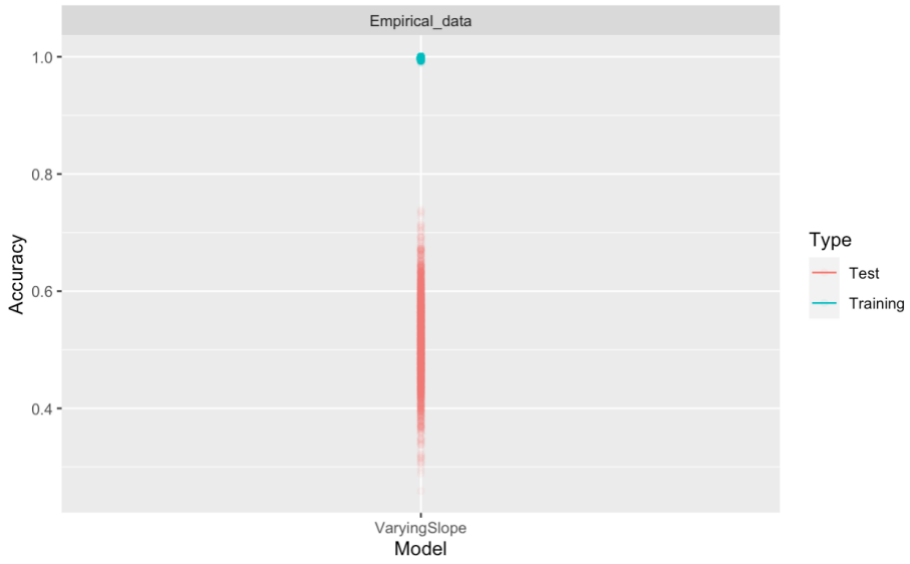
The output above indicates the global feature importance of the empirical data. Therefore, the top variables, that have the greatest importance, will be included in the model.

The data was split into two sets, the train set (80% of the data) and the test set (20 %). I have also attempted to balance the groups so that the same participant wouldn't appear in both sets, moreover, I attempted to roughly balance the gender and diagnosis among the test and training sets. The training data is scaled in the same way as the simulated data was. At first I haven't scaled the test set, however, it gave me errors about the levels (?), and after scaling errors disappeared, therefore I scaled the test data as well (but I feel like I shouldn't).

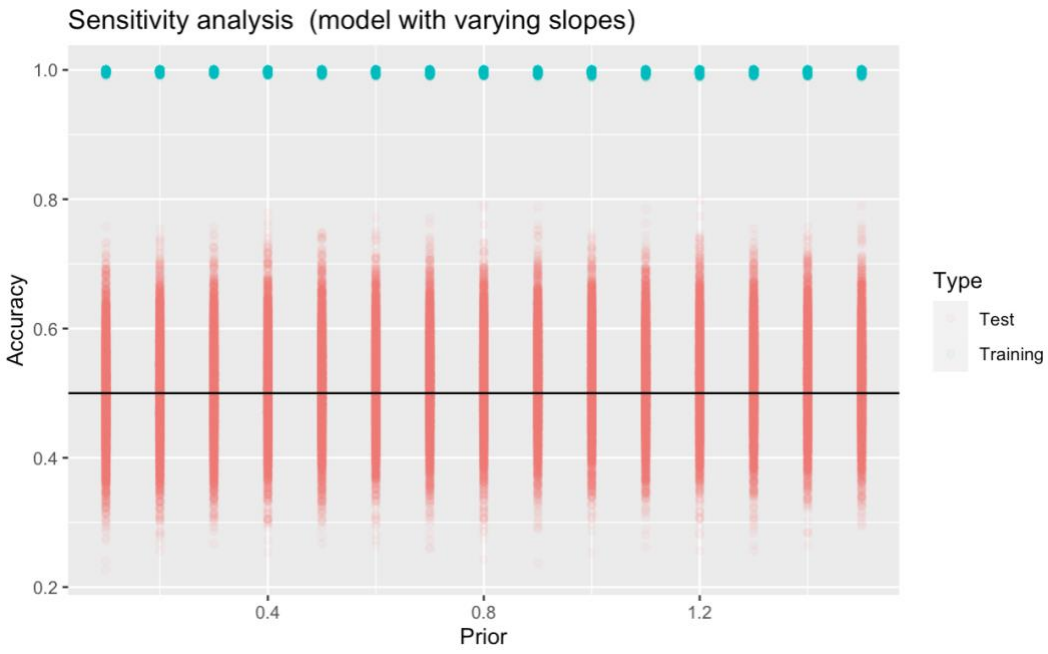
The first 13 variables, that showed the greatest importance, were included in the model. The posterior predictive check, however, provides results of having more controls in the data set than schizophrenic patients. In general, the whole empirical data set has less patients diagnosed as schizophrenics than controls, therefore the model might learn the patterns of acoustic measures of control group more confidently.



The average performance of the model was assessed as well. The model could classify the diagnosis of the training set almost perfectly (too perfectly I'd say?), however, had difficulty in classifying the test data set. In both cases, more schizophrenic patients were classified as controls, rather than controls classified as schizophrenics, which, I assume, makes the model more problematic.



The model seems to have a lot of uncertainty, the data of the test set gets classified poorly. For this reason, I have run the sensitivity analysis on the prior for the slope:



It showed that adjusting the prior for the slope would not affect the accuracy of prediction, therefore it seems to me that either I should run the prior check on the SD, or just change the model into a simpler one, or add more predictors (Gender?). But which way to choose, as prior sensitivity check for each of the model takes to me at least 3 hours to run, and feels very exhausting :D. I have tried to leave it to run during the night, but it just stops and begins again when I start using it.