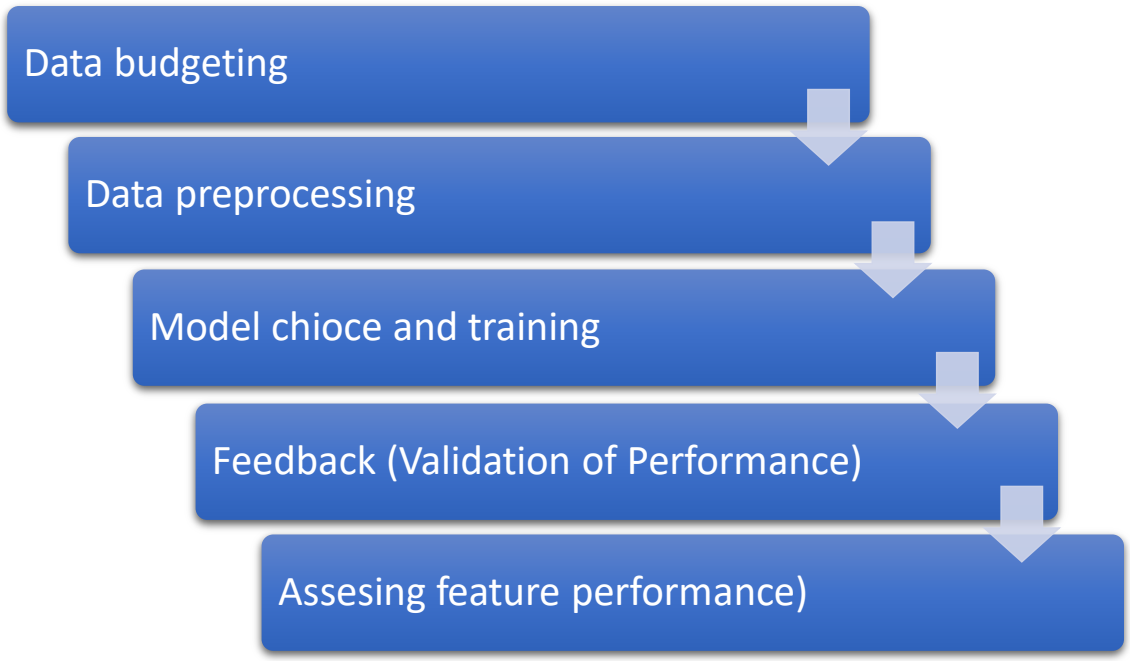
Ioana Luisa Forcas
au693527

Assignment 3

Question 1

Machine Learning Pipeline



Data budgeting

Training data makes up 80% of the total data, while testing data make up the remaining 20%. In order for the algorithm to learn from both the control group and the schizophrenic subjects, the split should respect the structure of the data, which means that participants from training data shouldn't emerge in testing data. The test set will be used to "check" if the algorithm could acquire and deduce the patterns, whereas the training set will be utilized for learning.

Data preprocessing

Scaling is necessary for the data (both simulated and empirical). Data scaling is done on the training dataset to ensure that population diversity information won't have an impact on how

well the algorithm performs on the test set. The test set is subsequently subjected to the same procedure (i.e., the mean and standard deviation).

Model choice and training

I first picked three models: model with fixed effects, varied intercepts, and different slopes in an effort to identify the model that could perform the best. The logistic regression approach is used to evaluate each model's quality (how effectively it can categorize an individual as being part of the control or schizophrenia group). In order to determine whether the priors need to be more cautious or wider, I also evaluated how the priors impact the models' effectiveness.

Assessment of performance

By examining the classification accuracy estimate and the types of errors the model makes, one may determine how effectively the model does classification. It's possible that the model incorrectly labels more "controls" as "schizophrenics," which could seem like a less serious error than labeling them the other way around.

Assessment of future importance

By examining the coefficients for each model and doing an analysis of the overall feature importance, the feature importance is determined. The findings show which characteristics are most important for categorizing the sample into the control and schizophrenia groups and are often employed by the model.

Question 2

Data simulation helps me better grasp the classification issue because I will be able to identify the components of simulated data and the information that is stored therein that is relevant for classifying people as normal or psychotic. Later, the machine learning pipeline is used as a "marker" to determine if the data patterns are inferred and whether participants can be appropriately categorized in accordance with predetermined metrics. Additionally, it will assist in determining which characteristics have the greatest influence on the categorization method. The outcomes of the simulation will enable a better understanding of the overall outcome of the empirical data and evaluation of its outcomes.

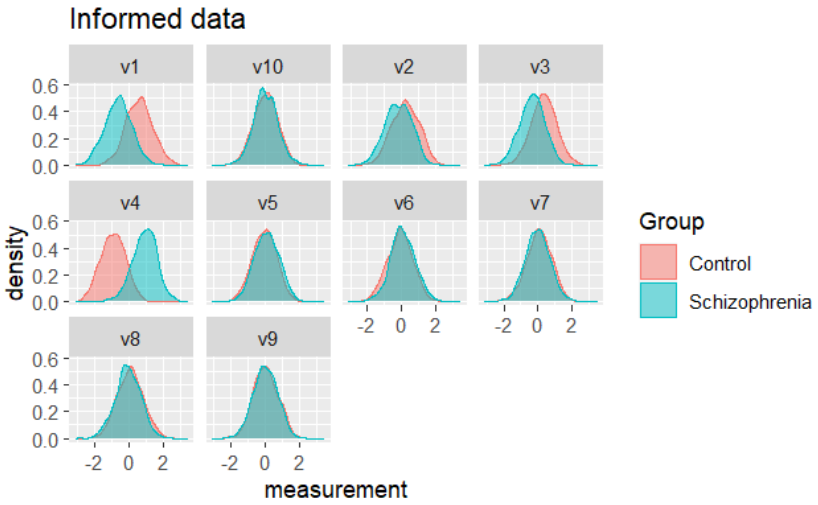
100 matched pairs of controls and schizophrenia patients were simulated in two datasets.

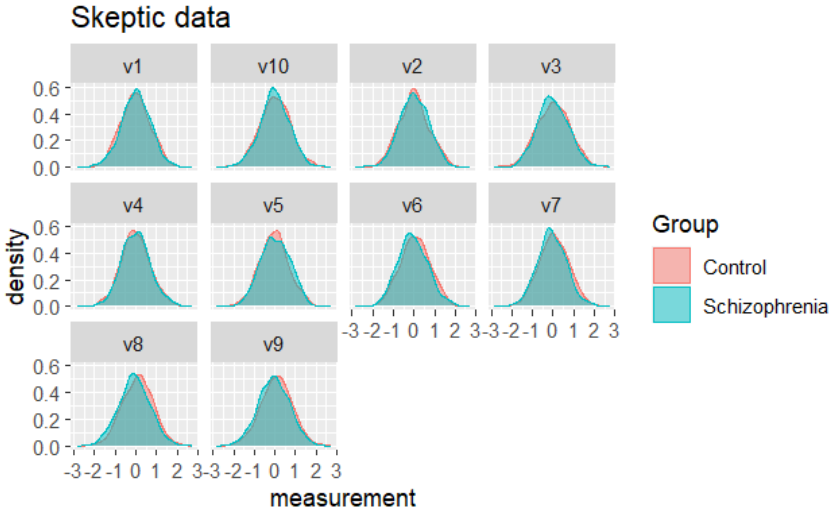
One sample ("skeptic") has 10 acoustic measurements and noise variables, whereas the other ("informed") contains 6 meta-analysis measurements and 4 random noise variables. Below is a description of the parameters:

Ioana Luisa Forcas
au693527

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

Below are charts of data from knowledgeable and skeptical sources. Measures v1–v4 show the biggest differences because they have the largest impact sizes, whereas measures v5–v6 are closer to 0 and appear to overlap more. There is no discernible change in the graphs either because the effect sizes in versions 7 through 10 are merely more noise and are the same for both data sets.





Ioana Luisa Forcas
au693527

Both informed and skeptical data sets are used for data budgeting. The remaining 20% of data is utilized as test data, with the remaining 80% being used as training data. Additionally, it was ensured that different participants wouldn't show up in the testing and training data sets.

The means and standard deviations of each characteristic were used to scale the measurements in the training data sets of the skeptical and informed groups (v1 – v10). The test data was treated using the same recipe.

I created three distinct models for informed and skeptic data sets independently in order to test which characteristics of the classification issue matter the most: one with fixed effects, one with variable intercepts, and the final one with 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)

Informed training data

When categorizing the patients into controls and schizophrenics, it appears that both fixed effects and changing intercept models perform equally on average (also have the same estimate of accuracy). With an accuracy of 0.995, the model with variable slope conducts classification more successfully.

Informed test data

The first two models behaved similarly (the same accuracy estimate). The model with constant effects, which categorized 7 schizophrenics as controls, performed better than the model with changing intercepts, which only classified 4 schizophrenics as controls. With different slopes, the model's reliability fell even further.

Skeptical training data

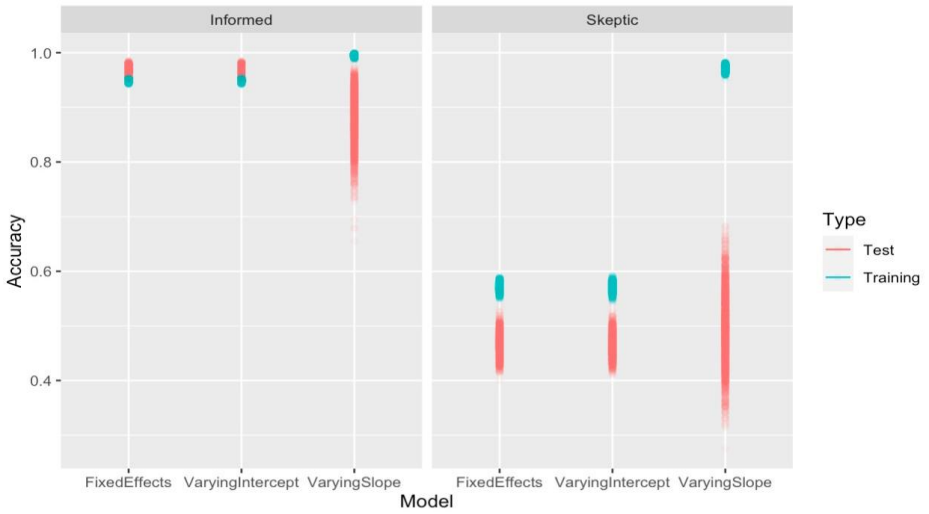
The model with variable slopes outperforms best, however it is not as accurate as the same model using informed training data. The first two models have about equal levels of accuracy.

Skeptical test data

The accuracy is just above the level of chance, demonstrating that models using skeptical data are unable to correctly categorize the diagnosis.

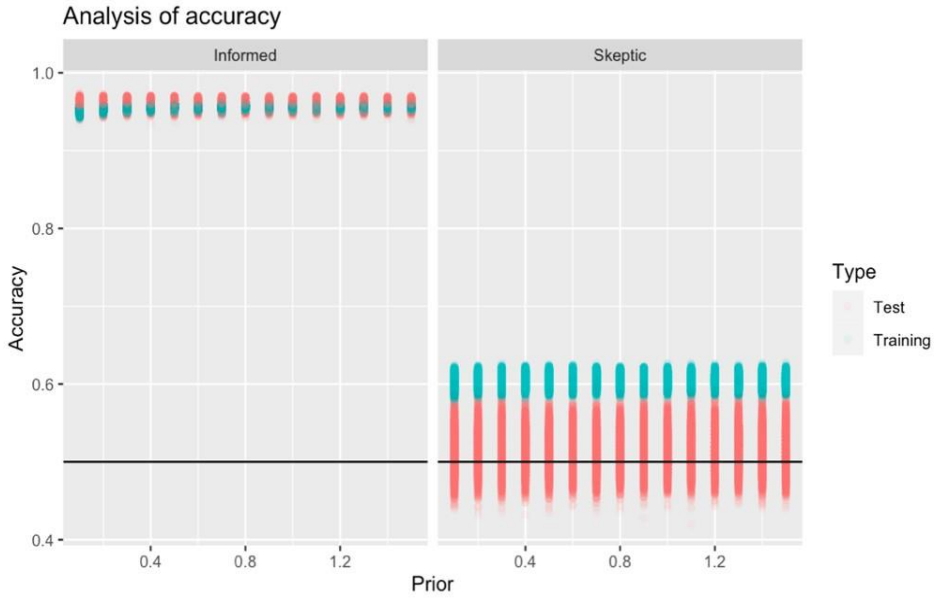
Ioana Luisa Forcas
au693527

Comparison when fitting the model:



Sensitivity Analysis

The graphic below shows how priors affect the model's ability to accurately divide individuals into groups according to diagnoses. Here, the fixed effects model's performance is recorded. Given that the ambiguity including both informed and skeptical data remains constant across all prior values, it appears that the prior has no effect on the precision of categorization.



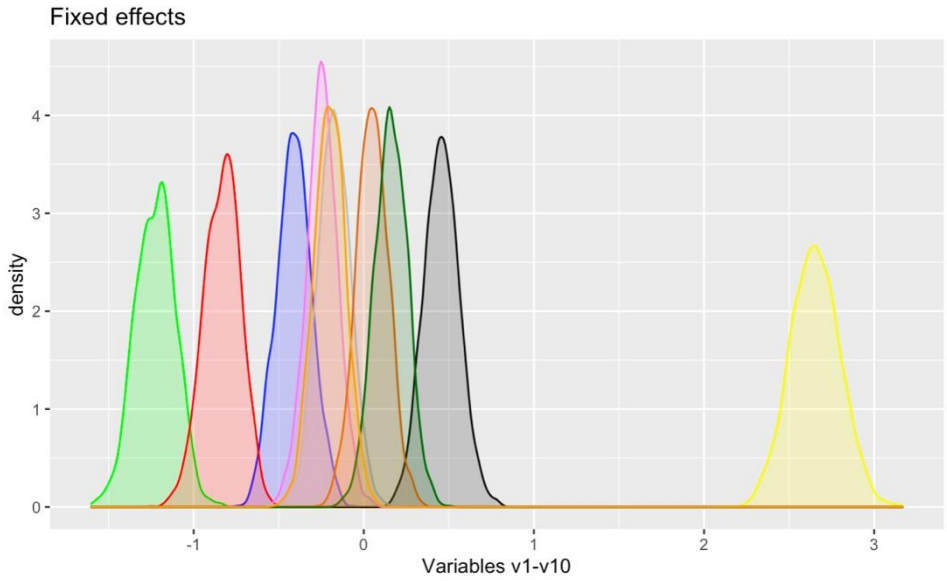
Ioana Luisa Forcas
au693527

Setting priors for the model with various slopes, however, yields quite varied results. Less conservative priors now lower the classification uncertainty (informed data) , but have no effect on the model uncertainty (skepticism data).

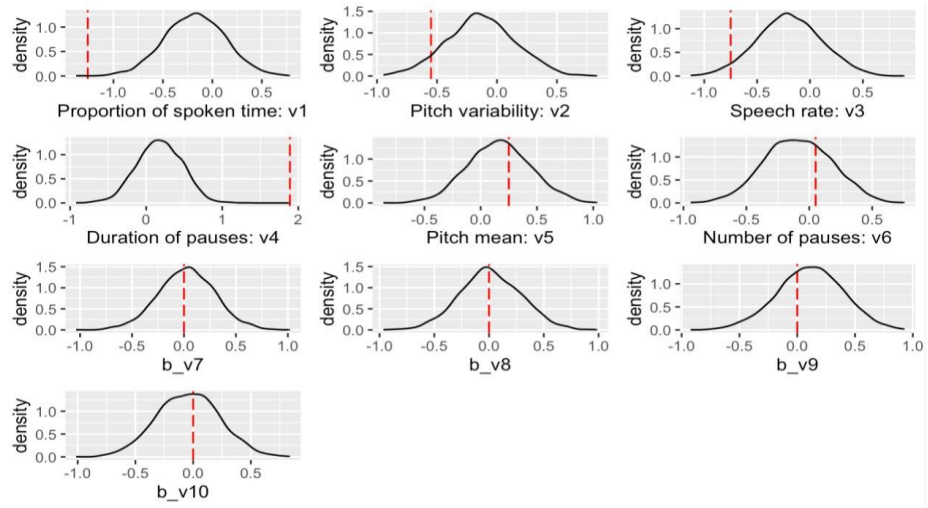
Using a thorough data set, I evaluated the feature relevance in each model. Below is a summary of each model's findings.

Model for fixed effects (and the same for varying Intercepts)

By examining the posterior distributions of each variable, it can be seen that the model heavily relies on the length of pauses (yellow), the percentage of spoken time (green), and the speech tempo (red).

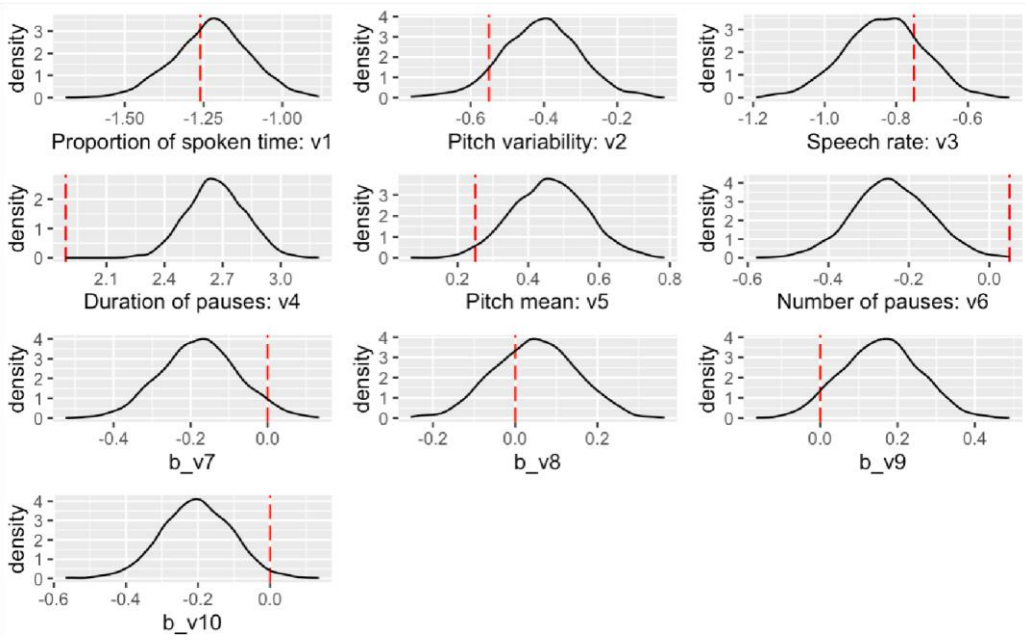


Density graphs for fixed effects:

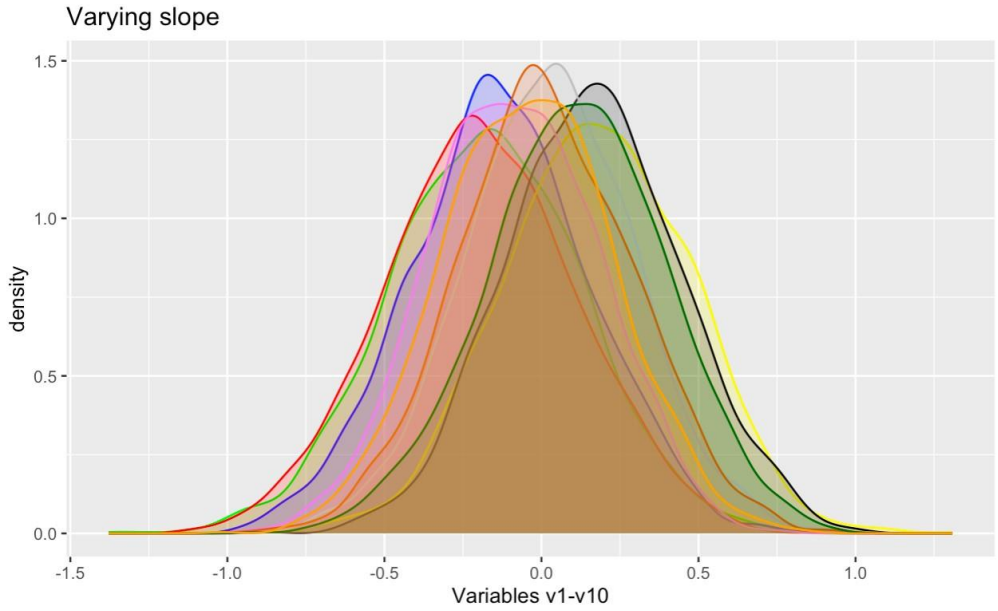


Ioana Luisa Forcas
au693527

Density graphs for varying Intercept:

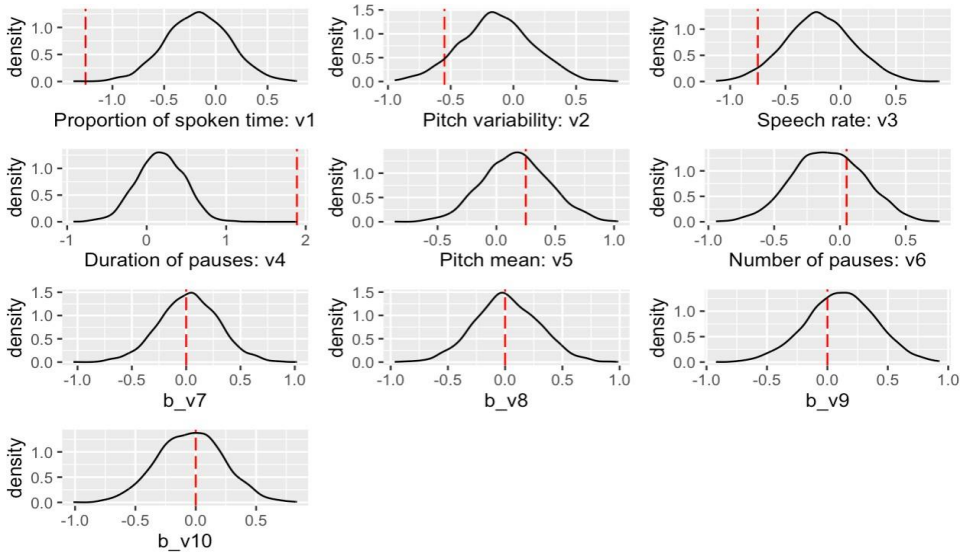


Varying Slopes:

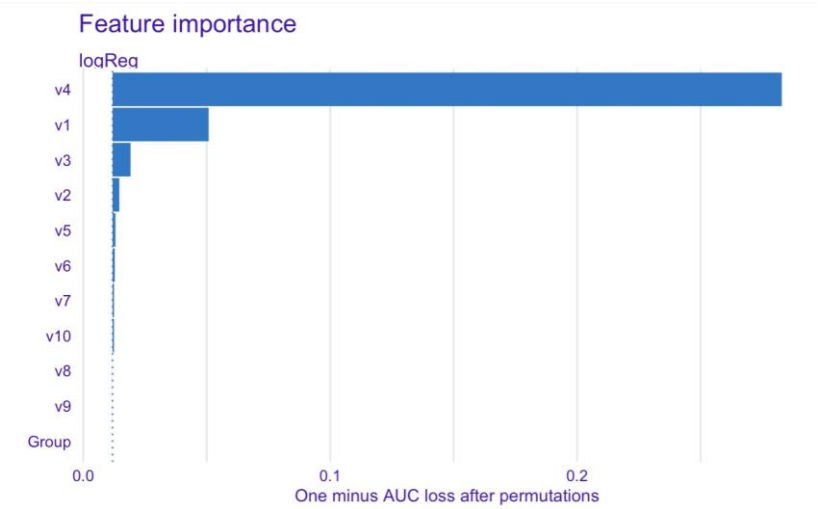


The distributions overlap significantly, making the length of the pauses (v4 - yellow) the key component of this model.

Ioana Luisa Forcas
au693527

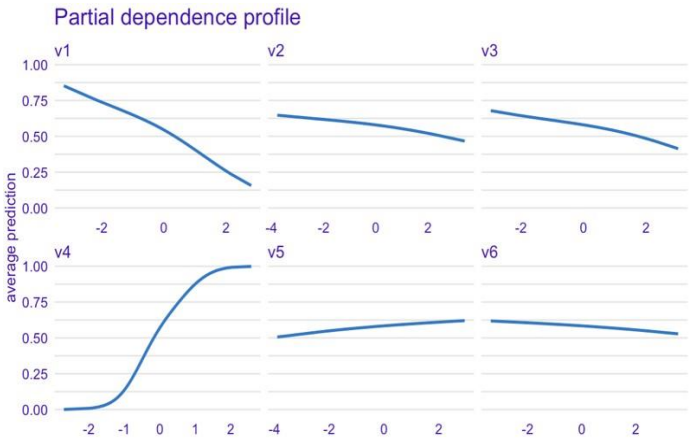


The findings show that, once again, v4 is the most significant feature that is employed, followed by the relevance of v1 and v3, when the logistic regression approach is used to evaluate the feature importance. Similar to the pattern with model coefficients previously discussed, the overall pattern of significance is present.



The next graph shows the relevance of the characteristics that the model is estimating. Therefore, the likelihood of being identified as having schizophrenia increases as the length of pauses increases (v4). Additionally, as the percentage of spoken time declines (v1), so does the likelihood of receiving a schizophrenia diagnosis. The fraction of spoken time and the measure of speech rate (v3) have similar patterns, although the latter is a more important metric. The "noise" measurements v7-v10 appear to be

important to the model in this plot, but since their influence on the model is zero and their plots are created on a lower scale than those of the measures v1-v6, they are just random noise that predicts the diagnosis only on a chance level.



Question 3

On empirical data, the machine learning process is used. The training set had 80% of the data, and the test set contained 20%. Additionally, I made an effort to fairly distribute the gender and diagnoses throughout the test and training sets, as well as to equalize the groups such that the same person wouldn't occur in both sets. There are 654 females and 868 males in the training data set, of whom 721 have schizophrenia and 801 have been classified as controls. There are 154 girls and 213 males in the test data set, of whom 179 are schizophrenia sufferers and 188 are controls. Consequently, it seems that the groups are even, with a larger proportion of men and controls in each.

The scale of the training data is the same as that of the simulated data. The test set was scaled using the same methodology as the training set.

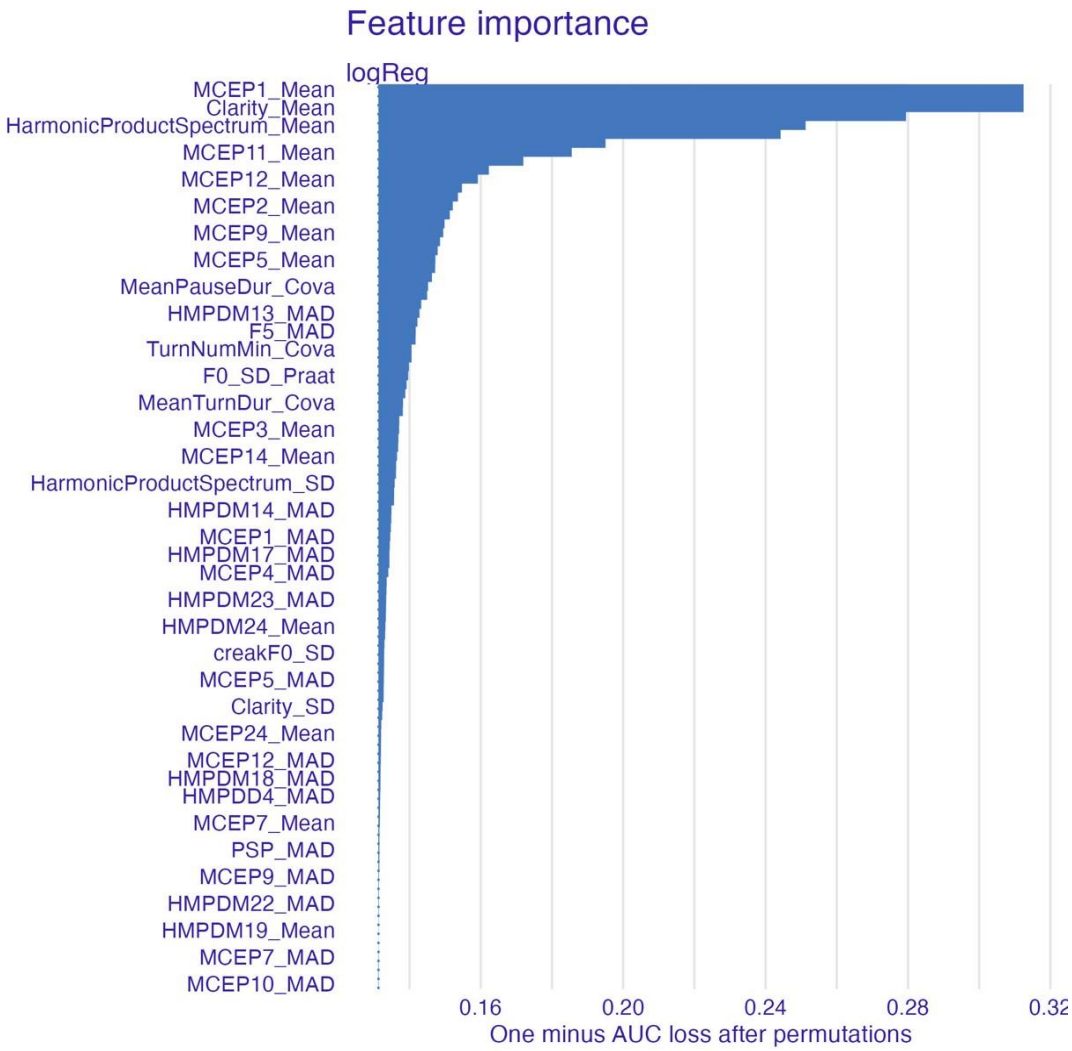
Since empirical data contains a much higher number of predictors than simulated data, I decided to train the model using the "tidymodels" library. Here, I've put up the clustering using logistic regression and random forest models. For the logistic regression model, the test set's accuracy in predicting the diagnosis was 0.664, whereas the training set's accuracy was 0.787.

Since the random forest model's accuracy is 0.73 on the test set and 1 on the training set, it is more accurate to classify data using it than logistic regression.

I chose to do a global feature relevance analysis on the logistic regression model to determine which characteristics are most crucial for categorization. I have taken off non-acoustic data like IDs, gender, and

language in order to achieve that. Additionally, the strongly associated characteristics (correlation > 0.7) were also taken out of the dataset.

Feature importance – Empirical data:



The predictors MCEP1 Mean, Clarity Mean, and HarmonicProductSpectrum_Mean are three factors that are employed the most when predicting the diagnosis, according to the results of the global feature significance analysis.

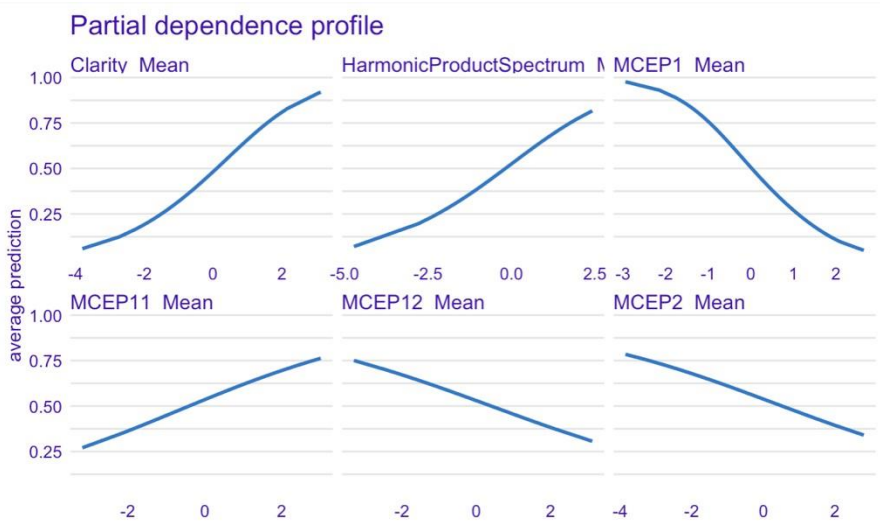
Ioana Luisa Forcas
au693527

The profile graphs above also show the estimated relevance that the model is estimating. According to the logistic regression technique, the likelihood of being labeled as SCZ (schizophrenic) increases as the mean of Clarity, Harmonic Product Spectrum, MCEP11, MCEP5, MCEP9, and MeanPauseDur Cova increases. The chance of being labeled as SCZ is decreased in all other scenarios illustrated above as a result of the decline in predictors.

The outcomes of the machine learning process reveal a great deal about the available empirical data. First and foremost, while performing data budgeting, it is important to understand the variables that make up the data and how they could influence the outcomes if balancing is not performed. In this instance, we ensure that there are no appreciable discrepancies between the two data sets by balancing the training and test data by gender and diagnosis.

When we do that, we learn more about the ratio of men to women in the data set as a general whole (as well as the diagnosis), which may make it easier for us to understand the prediction results. In this case, it is obvious that the empirical data contains more men and individuals who are categorized as "controls," so the algorithm may become more accurate when predicting diagnosis for those individuals.

When determining which metrics are most important in classification issues, it is useful to examine the value of global features. In real-world scenarios, the algorithms that effectively forecast the diagnosis (with the fewest mistakes feasible) may also be applied.



Ioana Luisa Forcas
au693527