Scientific Research Publishing

Open Journal of Statistics, 2020, 10, 1-9
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

# Predictive Models for Functional MRI Data

**Guenadie Nibbs[1], Peter Bajorski[2]**

[1]Dirección General de Impuestos Internos, Santo Domingo, Dominican Republic
[2]Rochester Institute of Technology, Rochester, USA
Email: g.nibbsc@gmail.com

**How to cite this paper:** Nibbs, G. and Bajorski, P. (2020) Predictive Models for Functional MRI Data. *Open Journal of Statistics*, **10**, 1-9.
https://doi.org/10.4236/ojs.2020.101001

**Received:** December 2, 2019
**Accepted:** January 6, 2020
**Published:** January 9, 2020



## Abstract

In this study, we analyze brain activity data describing functional magnetic resonance imaging (MRI) imaging of 820 subjects with each subject being scanned at 4 different times. This multiple scanning gives us an opportunity to observe the consistency of imaging characteristics within the subjects as compared to the variability across the subjects. The most consistent characteristics are then used for the purpose of predicting subjects' traits. We concentrate on four predictive methods (Regression, Logistic Regression, Linear Discriminant Analysis and Random Forest) in order to predict subjects' traits such as gender and age based on the brain activities observed between brain regions. Those predictions are done based on the adjusted communication activity among the brain regions, as assessed from 4 scans of each subject. Due to a large number of such communications among the 116 brain regions, we performed a preliminary selection of the most promising pairs of brain regions. Logistic Regression performed best in classifying the subject gender based on communication activity among the brain regions. The accuracy rate was 85.6 percent for an AIC step-wise selected Logistic Regression model. On the other hand, the Logistic Regression model maintaining the entire set of ranked predictor was capable of getting an 87.7 percent accuracy rate. It is interesting to point out that the model with the AIC selected features was better classifying males, whereas the complete ranked model was better classifying females. The Random Forest technique performed best for prediction of age (grouped within five categories as provided by the original data) with 48.8 percent accuracy rate. Any set of predictors between 200 and 1600 was presenting similar rates of accuracy.

## Keywords

Functional Magnetic Resonance Imaging, Regression, Logistic Regression, Linear Discriminant Analysis, Random Forest

## 1. Background Information

Understanding the human brain has been one of the most important topics studied by neuroscience. This field has come up with different imaging theories that are used to quantify the properties of brain networks and their components. The brain has been modeled as a complex network system under the premise that neurons make up an interconnected structure of the nervous system.

The development of new techniques for image acquisition has made possible the improvement of extracting quality information of brain activity. By utilizing functional magnetic resonance imaging (fMRI), it is possible to measure brain activity, based on changes of the oxygen level in the bloodstream over time.

This paper investigates the relationships among different brain regions and how the nature of those relationships might be a predictor of the subject's characteristics. The analysis will be based on information gathered from the fMIR of 820 subjects, performed under similar conditions (Resting-State). Various statistical techniques are used to predict subject traits based on brain connectivity.

## 2. Regions of Interest (ROIs)

The data coming from fMRI images is given by voxels. Depending on the objective of every researcher, the characteristics of a node could vary. For this reason, nodes should represent, meaningfully and accurately, the elements to be investigated in the system. In a brain network, theoretically, the most accurate representation of a node would be an individual neuron, having all synapses representing its links. The issue with this approach is that existing technology can only account for areas over 1 mm, while neurons sizes are around 0.004 mm [1]. Furthermore, all signals coming from those nodes are weaker in comparison with other alternative representations and hence, harder to interpret because they contain more noise.

The minimum size a node could take is that of a voxel (1 mm). The issue with this representation is that there are around four million of voxels in a human brain image, each one of them with around eight thousand synapses. Applying computational procedures or even recording this amount of data would represent a huge challenge for any researcher. Because of these technical limitations, the bigger the amount of neurons or voxels used to represent a single node, the easier it is to perform computational analyses on them. In this grouped representation, all interacting neurons and synapses within that given space represent a singular node in the brain. The challenge with this representation comes with the fact that because nodes can be built freely in terms of size and location, the selection of these features needs to be done carefully depending on the researcher's objective. At the moment of selecting the spatial area of each node, it is necessary that the area shares similar features.

The Regions of Interest represent the variety of ranges in the number of nodes and their locations used to create a bigger new one; and the parcellation scheme they have to use to maintain an accurate interpretation of the results. This means

that depending on the characteristics of the ROIs, interpretation of the results could differ.

## 3. Parcellation Scheme in MRIs and fMRIs

Parcellation schemes split the spatial brain area into a set of non-overlapping regions that present homogeneity with respect to their components. Regions of interest not always are at the same level of an individual voxel, but they could be at the level of a set of them. This situation gives place to the existence of two different modalities that could be categorized: single voxel-based and aggregates voxels-based. In the single voxel-based modality, an individual voxel will represent a single node [2]. This modality has been seen as one of the best representations of relationships within the system.

The aggregates voxels-based modality takes two forms, Multi-Voxel analysis and Brain Atlases. The multi-voxel analysis allows the researcher to define any structure of interest, while brain atlases provide a pre-defined set of regions with a certain base on the brain structure. Because these methods are based on voxels combinations to create the main unit (node) as a bigger entity of the system, blood oxygen level dependence signals need to be average within the ROIs.

The Automate Anatomical Labeling (AAL) has been used to identify 116 (ROIs) brain regions, 58 on each hemisphere of the brain. The fMRIs time-series within each region are averaged, and the new calculated time-series is used to describe the brain activity in those brain regions.

## 4. Statistical Analysis

### 4.1. Exploratory Data Analysis

The data used to perform the following analysis comes from the Human Connectome Project. The main objective of this project is to build a network map that provides a better understanding of anatomical and functional connectivity of the human brain [3]. The data consist of the extracted information from a neuroimaging sequences coming from fMRIs, which provide a measure of brain activity based on its functions over time. A connectivity matrix was calculated for each subject with the averaged time series based on the Automate Anatomical Labeling's Regions of Interest. The data used here consists of four fMRI scans for each of 820 subjects. Each scan is described by 116 time series representing 116 ROIs, and each time series.

Relationships between brain's regions are described by correlations between the respective time series [4]. The four correlation coefficients calculated based on the four scans for the same subject can be regarded as characteristic for that subject in terms of interaction between the two brain regions. We attempt to predict subject's traits based on those correlations between regions.

### 4.2. Features Selection and Summarization

We first determine the correlation value for each combination of two regions.

We then identify which of those regions have the best subject-to-subject consistency. Those are the correlations that are not much different between scans. Analysis of variance was used to assess consistency correlations within the four scans of each subject. To this end, a one-way ANOVA was set up with correlations as the dependent variable (response) and the subjects as the independent variable (predictor or input). Independent variable (subjects) was set as a random factor.

To get an accurate interpretation of correlations consistency, ANOVA should not be applied to the whole dataset at once. Instead, the process needs to be applied individually over each region combination. These results are then used to extract the variance components of the random effects and build a matrix using the variability coming from residuals.

This matrix plot in Figure 1 shows the level of variability coming from the error term in percent of the total variability. We desire low error percentages, which signify high consistency among the four scans of the same subject. The low values are represented in the plot with darker colors.

### 4.3. Predictive Models with Built-In Cross-Validation

The next step is to develop predictive models for age and gender of subjects. The correlations identified earlier as being the most consistent within scans for the same subject will be used as predictors. This process will involve three steps: features summarization, model building with built-in cross-validation and accuracy assessment.

The first subject characteristic selected as the response variable is Gender, with two categorical values: male and female. The values of the input variables are calculated as the square root of averaged variances from the four scans for a given subject. Because the response is a categorical variable, the problem becomes a classification problem. The error percentages previously calculated will serve as reference to include meaningful features in the model. The error percentages were ordered from lowest to highest, and a feasible number of 120 predictors was chosen from the top of the list.

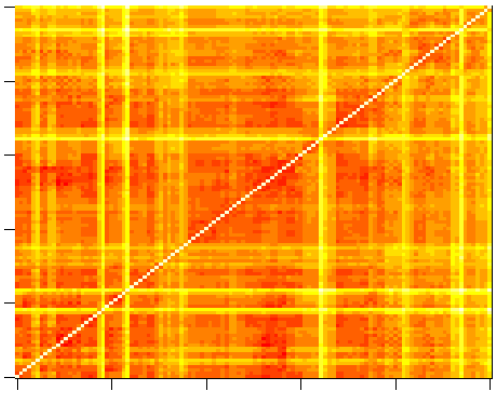

Figure 1. The matrix plot of the error term variability in percent of the total. The low values are represented in the plot with darker colors.

Model 1: Logistic Regression with AIC Selection-Gender: After the selection procedure, the resulting model possesses 40 predictors. The AIC model is capable of successfully classifying 384 females as females, and 318 males as males. On the other side, it incorrectly classified 47 females as males, and 70 males as females. Based on these results, the model performed predictions with 85.71 percent of accuracy. We then validated how the model performed in 10-fold cross-validation.

While the model without cross-validation and optimal threshold was able to predict with an accuracy rate of 86.20%, the cross-validated model got an 85.59% accuracy rate.

Model 2: Logistic Regression with Ranked Predictors-Gender: The model obtained using logistic regression contains 29 variables with p-values lower than 0.05 for statistical significance, intercept included. The model performed predictions with an 85.71 percent of accuracy. While the model without cross-validation and optimal threshold was able to predict with an accuracy rate of 88.64%, the cross-validated model got an 87.67% accuracy rate.

Model 3: Linear Discriminant Analysis with Predictors-Gender: The next statistical technique proposed for this classification problem was Linear Discriminant Analysis. This method allows characterizing two or more classes of objects based on means and variances, whose results must be used as a linear classifier. The LDA model was capable of successfully classified 402 females as females, and 321 males as males. On the other side, it incorrectly classified 52 females as males, and 44 males as females. Based on these results, the model performed predictions with 88.28 percent of accuracy. The accuracy rate for the trained model and the best prior is almost the same. On the other hand, the cross-validated model using best prior, is presenting a lower accuracy rate of 49.08% (**Figure 2**).

Model 4: Random Forest with Ranked Predictors-Gender: The last statistical technique to be applied in this classification problem will be Random Forest. This is a more general technique that uses a multitude of decision trees to determine which class is the best for the object to be classified. The accuracy rate corresponding to the random forest technique, when using 120 ranked predictors with the mtry parameter constant, is 72.80 percent.

Using a dynamic value for mtry, it shows that the best configuration for this set of ranked predictors correspond to best 68, where the accuracy rate ends up at 73.02 percent.

Model 5: Regression with Ranked Predictors-Age: The second subject characteristic selected as the response variable was Age. Similar to the previous models, the regression analysis will be performed using the ranked predictors, thus the model accuracy could be compared with the others at the same level. The model obtained using regression analysis contains 12 variables with p-values lower than 0.05 for statistical significance, intercept included. While the r-squared has a value of 0.2881, the adjusted r-squared has a lower value of

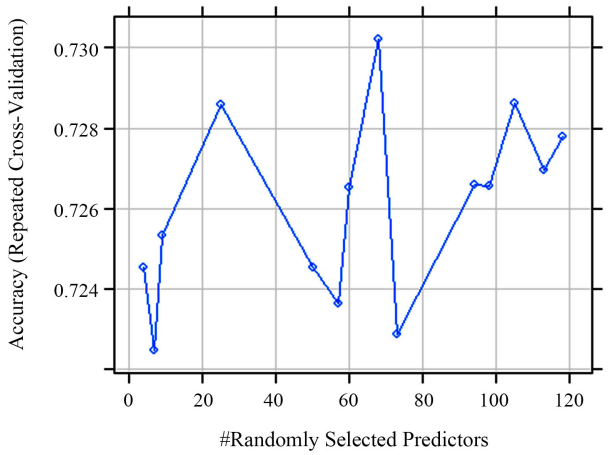


**Figure 2.** Graphical representation of the accuracy given different numbers of predictors for Gender.

0.1658. Based on these results, the model performed predictions with 9.1 percent of accuracy.

Model 6: Random Forest with Ranked Predictors-Age: We now apply Random Forest to the same set of ranked predictor and evaluate the performance improvement. The tuning parameter "mtry" was held constant at a value of 11. The Random Forest with constant mtry got similar results to the previous model. The r-squared value is also close to 9 percent.

Model 7: Random Forest with Ranked Predictors-Age as Categorical: Initially the ordinal response variable was transformed to numerical type as a way to avoid losing order information. This process was done taking the mid-point of the range of every category. Because the model did not perform well, the same statistical technique is now applied over the same set of values but using a classification perspective. Random Forest allows performing models for both, prediction and classification cases. The first configuration will be maintaining mtry constant value of 11 over the whole procedure.

The model was able to classify categories with an accuracy of 47 percent. Now let us see if there is a change coming from setting mtry dynamic. Accuracy was used to select the optimal model using the largest value. The final value used for the model was mtry = 113.

When configuring the mtry value as dynamic, the best configuration for this set of ranked predictors corresponds to 113 selected predictors, where the accuracy rate ends up at 48.04 percent (Figure 3).

## 5. Accuracy Assessment and Recommendations

Having applied four different statistical methods (Regression, Logistic Regression, Linear Discriminant, Random Forest) to classify/predict two relevant subject's traits, it is possible to make assessments on how these models performed based on the accuracy rate obtained with each method. For contrast purposes, all models were performed using the same set of ranked predictors, which makes

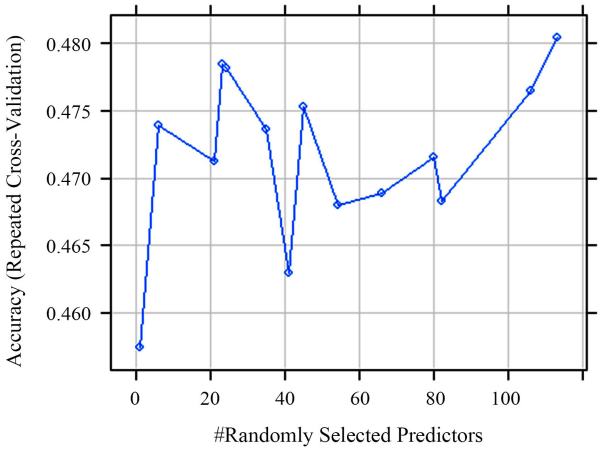


**Figure 3.** Graphical representation of the accuracy given different numbers of predictors for Age.

possible to determine the best choice using a similar amount of computational resources. The following table shows a summary of the accuracy measurement for each technique at every level of optimization.

Both prediction and classification analysis get different accuracy measurements at each level of the process. The standard level corresponds to the training of the model using the entire dataset and using the same values to predict. The second level corresponds to the same standard process but adding an optimization technique to determine the best threshold. The last level represents a cross-validation procedure utilizing the optimal threshold.

The motivation for using cross-validation is to avoid overfitting. Without cross-validation, the accuracy measure only tells how the model performs in that specific dataset. The main interest in this case is having an accuracy measure that could represent the correctness of the model for any new dataset of this type. For this reason, the goodness of the model will be evaluated based on cross-validated results (**Table 1**).

Selecting 120 ranked predictors to perform each statistical technique was needed in order to balance between getting an adequate accuracy rate, managing viable computational times, and avoiding irrelevant predictors. The linear discriminant technique had a good performance using the optimal prior, but it fell down in the cross-validation procedure going from 88.28 to 49.08 percent accuracy rate. For this reason, this was the first discarded technique of the three used to model gender. Random Forest also performed well using mtry set constant and little bit better when the parameter was dynamic. It went from 72.80 to 73.02 percent accuracy rate. It was the most robust technique, allowing to model gender when using over a thousand predictors. The results with more than 200 predictors were not included here, because they did not affect much the accuracy rate (about 1 percent better, but 5 times slower on the computational side). Although the Random Forest model had a good performance and the best robustness, it was discarded because the last two models outperformed its results.

**Table 1.** Table of the accuracy measurement for each technique at every level of optimization.

| Model Configuration | | | Accuracy | | | |
|---|---|---|---|---|---|---|
| Predictors | Response | Statistical Technique | AIC/R-Sq | Standard | Optimal/Mtry | CV |
| 120 | Gender | AIC Logistic Regression | 616.79 | 85.71 | 86.2 | 85.59 |
| 120 | Gender | Logistic Regression | 699.53 | 87.79 | 88.64 | 87.67 |
| 120 | Gender | Linear Discriminant | NA | 88.28 | 88.28 | 49.08 |
| 120 | Gender | Random Forest | NA | NA | 73.02 | 72.8 |
| 120 | Age | Regression | 9.1 | NA | NA | NA |
| 120 | Age | Random Forest | 9.75 | NA | NA | NA |
| 120 | Age | Random Forest | NA | NA | 48.4 | 46.99 |

Logistic Regression performed best in classifying the subject gender based on functional connectivity. The AIC Logistic Regression model was capable of getting an 85.6 percent accuracy rate. Alternatively, the Logistic Regression model maintaining the entire set of ranked predictor was capable of getting an 87.7 percent accuracy rate. It is interesting to point out that the model with the AIC features was better in classifying males, whereas the complete ranked model was better in classifying females.

Even though the Logistic Regression technique was not as robust as the Random Forest, it was able to get better accuracy rates after cross-validation. Moreover, because this type of model is based purely on linear relationships, is easier to explain and it can be easier implemented by other researchers with low or no expertise in statistical analysis.

When considering Age as the response variable, the first technique, corresponding to regression analysis, failed trying to capture the pattern to predict the subject's age. This variable was given as an ordinal type level of measurement. The first approach consisted of converting each category to continuous in order to avoid losing information coming from the order. In the same way, Random Forest was performed using the same specification and also failed, getting an r-squared of 9.75 and 9.10 for the regression technique.

The results improved when the variable was treated as a nominal type with five categories. The Random Forest technique using mtry dynamic was capable of getting 48.80 percent accuracy rate. Any set of predictors between 200 and 1600 was presenting similar rates of accuracy.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Finn, E.S., *et al.* (2015) Functional Connectome Fingerprinting: Identifying Indi-

viduals Using Patterns of Brain Connectivity. *Nature Neuroscience*, **18**, 1664-1671. https://doi.org/10.1038/nn.4135

[2] Vul, E., Harris, C., Winkielman, P. and Pashler, H. (2009) Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, **4**, 274. https://doi.org/10.1111/j.1745-6924.2009.01125.x

[3] Van Essen, D.C., *et al.* (2013) The WU-Minn Human Connectome Project: An Overview. *Neuroimage*, **80**, 62-79. https://doi.org/10.1016/j.neuroimage.2013.05.041

[4] Gabrieli, J.D., Ghosh, S.S. and Whitfield-Gabrieli, S. (2015) Prediction as a Humanitarian and Pragmatic Contribution from Human Cognitive Neuroscience. *Neuron*, **85**, 11-26. https://doi.org/10.1016/j.neuron.2014.10.047