Predicting Spatial abilities using psychological features

A STATISTICAL LEARNING PROJECT

Grimaldi Francesco
Universita degli Studi di Padova
Master Degree in Data Science

June 2019

Abstract

The aim of this project is to identify what are the personality traits and the individual self-reported wayfinding inclinations (i.e what is a person behaviour when he/she must deal with visuo-spatial tasks) which can predict the best the actual visuo-spatial abilities, where in our case are two.

We tried to identify the best and smallest subset of features, starting from the initial 27 features and by using different statistical methods for binary classification problems such Generalized Linear Model (GLM), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and K-Nearest Neighbours (KNN). The results show that spatial abilities cannot be predicted with high accuracy but can reach discreet accuracy with only a very limited number of features such as Gender, some personality traits regarding the emotional stability of a person and how much a person tends to use the cardinal points to orientate him/herself during a visuo-spatial task. The accuracy of a polynomial GLM with degree two reached 0.739, while a GLM of degree one and with only two features, Gender and Emotion Control, scored a 0.676 accuracy.

Contents

Chapter 1

Dataset's description

1.1 Instances description

The dataset is made by 222 instances that in our case are undergraduates students at the School of Psychology at the University of Padua (Italy). Their mean age is 20.52 with a standard deviation of 1.36. Furthermore is important to note that of the 222 subjects only 80 identify as male while the others 142 identify as female.

1.2 Features description

The dataset collect a total of 29 variables for each participant. Those 29 variables can be divided in 5 categories:
1. *Gender*: a categorical variable where 1 represent male subject while 2 represent female subject.
2. *Personality traits*: fifteen numerical variables that represent the main personality dimensions.
3. *Navigation Aid Questions* (NAD or QLF): three numerical variables which try to measure the behaviour of navigation of subjects.
4. *Self-reported wayfinding inclinations*: eight numerical variables, with six that try to measure the self-perception of emotions and behaviour of the participants when dealing with visuo-spatial tasks and the remaining two are two latent variables of those previous six dimensions.
5.*Spatial Abilities*: two numerical variables that measure trough two test the actual visuo-spatial capabilities of a person.

1.2.1 Personality traits

This part is composed by 15 numerical features coming from the *Big Five Questionnaire* (BFQ; Caprara, Barbanelli, Borgogni, Perugini, 2008).
The BFQ is composed by 134 items on a Likert Scale from 1 to 5, referring to 5 traits, with 2 facets for trait. The 15 features represent the 5 traits (Extroversion, Agreeableness, Consciousness, Emotional Stability and Openness) and the 10 facets (Dynamism, Dominance, Cordiality, Cooperativeness, Scrupulosity, Perseverance, Pulse Control, Emotion Control, Experience Opening and Culture Opening).
It's important to point out that even if the items are in categorical scale, it has been decided to consider the score of the facets as numerical since the values ranges from 12 to 65 and are in a ordinal scale.

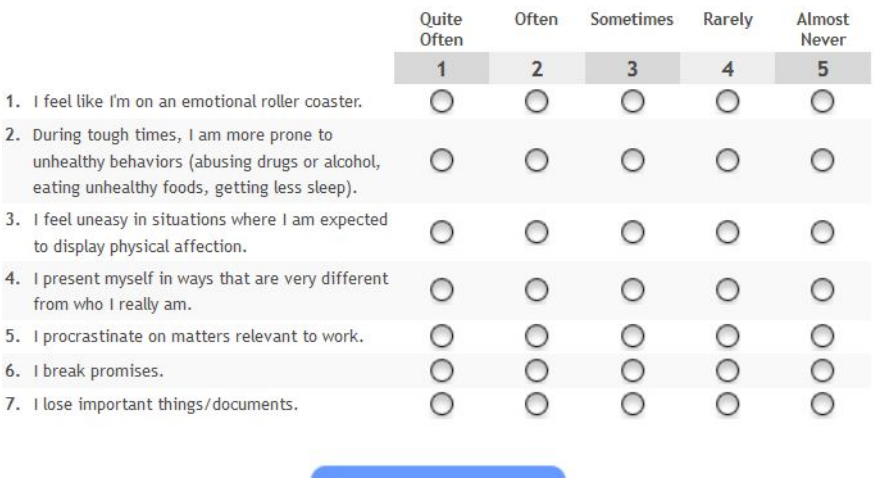

Fig 1.1. Few examples of items of Big Five Questionnaire

1.2.2 Navigation Aid Questions

Composed by three items with a Likert Scale from 1 to 6 (I disagree/I agree), the NAD (adapted by Munzer, Zimmer, Schwalm, Baus, Aslan, 2006) investigates what a person use during navigation tasks.
The first item ask if a person uses a map, the second item ask if a person uses a GPS and the third one ask if a person tends to ask for verbal indication.

1.2.3 Self-reported wayfinding inclinations

This part of features is composed by six numerical features coming from three different questionnaires and two numerical features coming from a previous factorial analysis of

those last six features (Meneghetti, Grimaldi, Nucci, Pazzaglia, 2019). Those six dimensions are:

1. *Space Anxiety Scale (QAS)* from its homonym questionnaires (SAS; adapted from Lawton, 1994; De Beni et al., 2014). It is composed by 8 items assessing the degree of space-related anxiety experienced in everyday spatial tasks.
2. *QACOknown*: coming from the *Attitudes to Orientation Tasks scale* (AtOT; De Beni, Meneghetti, Fiore, Gava, Borella, 2014). It asses the preference for navigating in well-known places.
3. *QACOexploration*: coming from AtOT, it asses the pleasure in exploring new places.
4. *Sense of Direction*: coming from the *Self-reported wayfinding inclinations Sense of Direction and Spatial Representation questionnaire* (SDSR; Pazzaglia Meneghetti, 2017), it asses your confidence in your ability to not to get lost.
5. *Cardinal Points*: coming from SDSR, it asses your confidence in using the cardinal points to orientate yourself.
6. *Landmark and Route Mode*: coming from SDSR, it asses your preference to use less complicated strategy to orientate yourself.

The last two dimensions, as previously said, come from a factorial analysis done to those six dimensions and from the same dataset. The resulting two factors are called *Positive Factor* and *Negative Factor* and are the sum of the standardized components weighted by their factor analysis's loadings.

1.2.4 Spatial Abilities

It's composed by two numerical features, which will be our dependant variables for the classification models.

The first feature is the Mental Rotations Test (MRT; adapted from Vandenberg, Kuse, 1978; De Beni et al., 2014). It consists of 10 items each one requiring in identifying two of four abstract 3D objects matching a target object but in a rotated position. The maximum time given is 5 minutes. For each item a point is given if both alternatives was detected (Maximums score: 10).

The second feature is the Point Perspective Taking task (PTT; adapted from Kozhevnikov and Hegarty, 2001; De Beni et al., 2014). It consists of 6 items each one requiring to imagine standing at one object in a 7-object configuration, facing another, and drawing the direction of a third on a circle. The maximum time given is 5 minutes. It is calculated the degree of difference between the angle individuated and the correct one and the sum

of the degree of differences is made (i.e. higher score high number of errors).

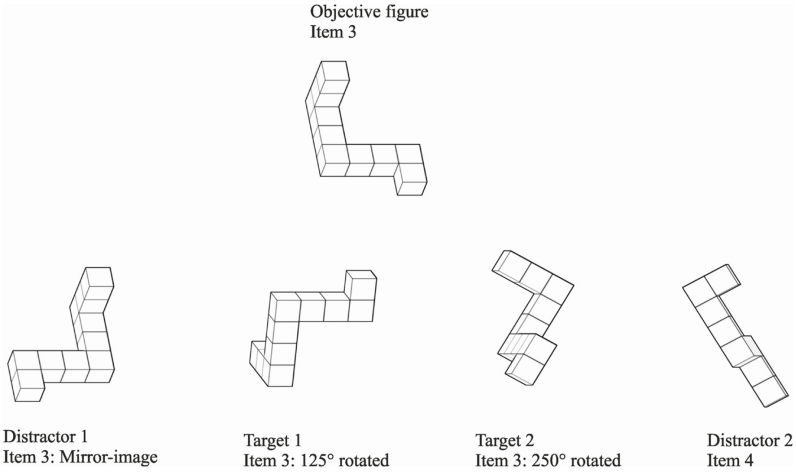

Fig 1.2. An example of an item in the Mental Rotation Task

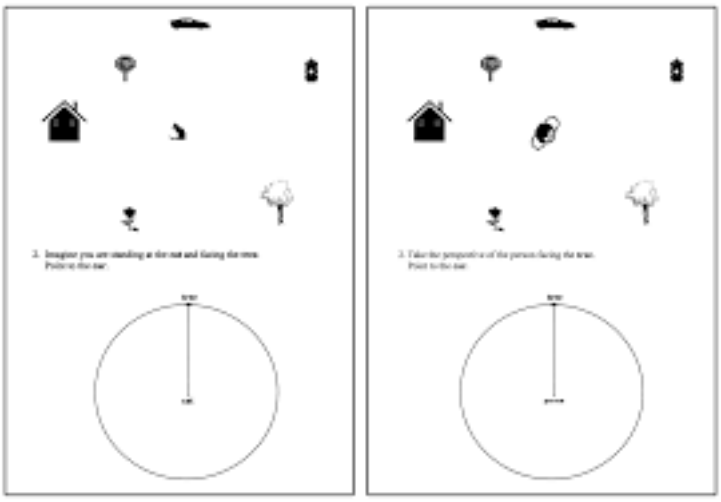

Fig 1.3. An example of an item in the Point Perspective Taking Task

1.3 Recap of the dataset

In this chapter we have described our dataset. A dataset of 222 person with 29 features from each person. Those features can be divided in Gender, personality traits (BFQ) , navigation aid questions (NAQ), self-reported wayfinding inclinations and visuo-spatial abilities. Furthermore gender and NAQ are treated as categorical variables where the other are treated as numerical.

Chapter 2

Preliminary analysis

2.1 Data cleaning

The dataset (format .csv) when imported in R contained much more columns then the actual features, those extra-columns were filled with NA and so they were cut off from the dataset.
Another operation that had been done was to check if other NA values were present in the dataset. It has been found that were present 7 NA values, all in the PTT features, since the low numbers of instances, it has been decided not to eliminate those instances but to keep them by inserting the mean of the PTT scores in those rows.

2.2 From numerical to categorical

After the cleaning process a brief look to the descriptive statistics and the data type of the features has been taken. Since all the features were treated as numeric by R, a conversion function from numeric to factor/categorical has been applied to the Gender.

2.3 Basic data visualizations

Some data visualization has been performed in order to check the distribution (histograms) and the normality of the data (qqnorm).

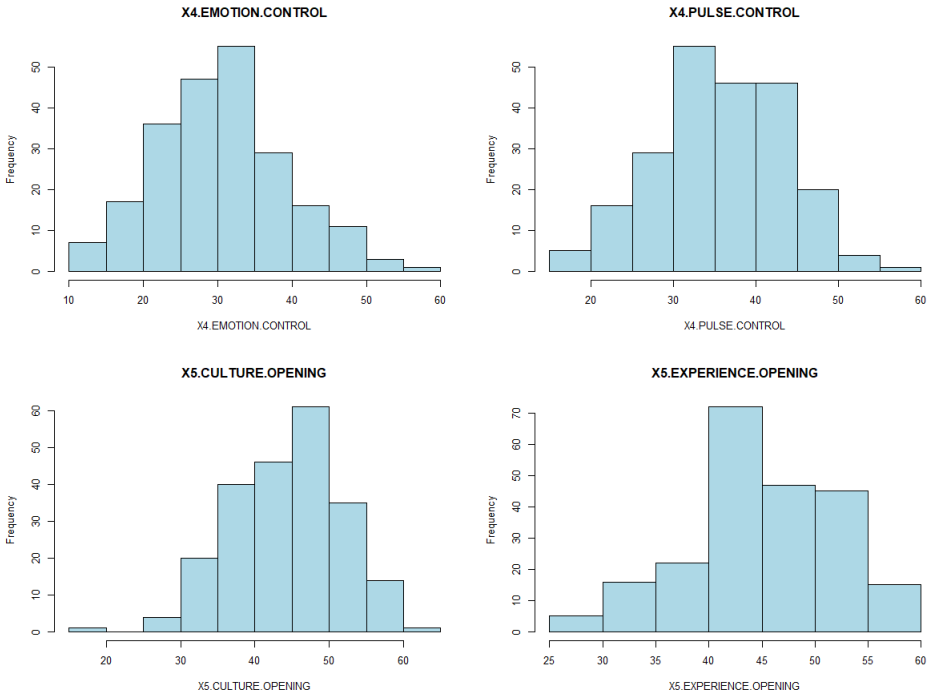

Fig 2.1. Distributions of some BFQ facets

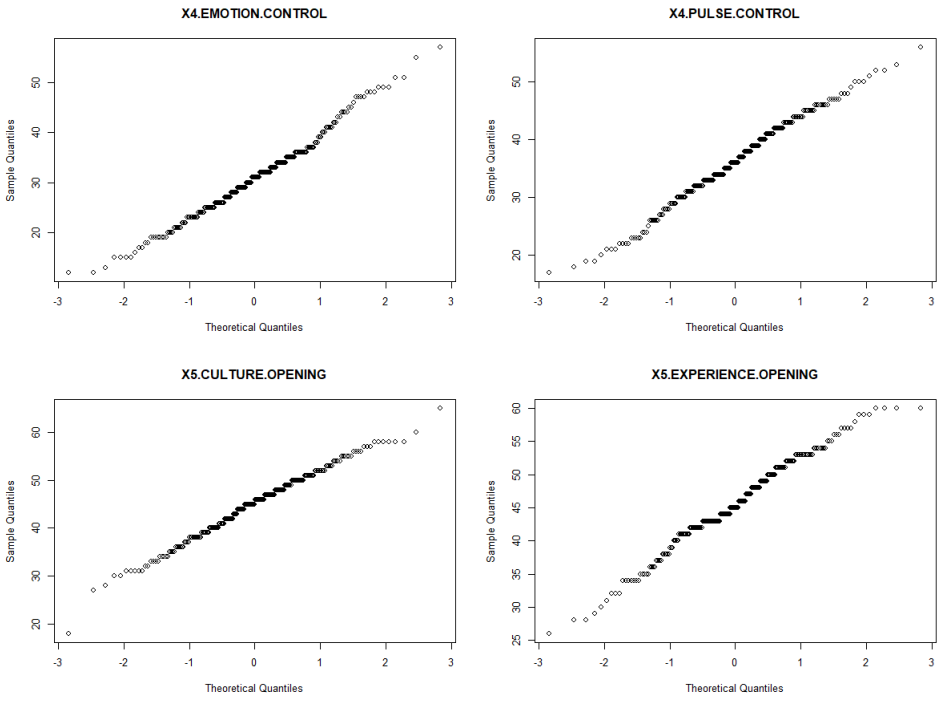

Fig 2.2. Normality of the data of some BFQ facets

2.4 PTT distribution: a special case

Since the PTT distribution seemed not normal and the sample quantiles against the theoretical ones seemed to form more than an exponential line than a straight one, a log transformation has been applied to the PTT. After this transformation the data seemed to appear much more normal

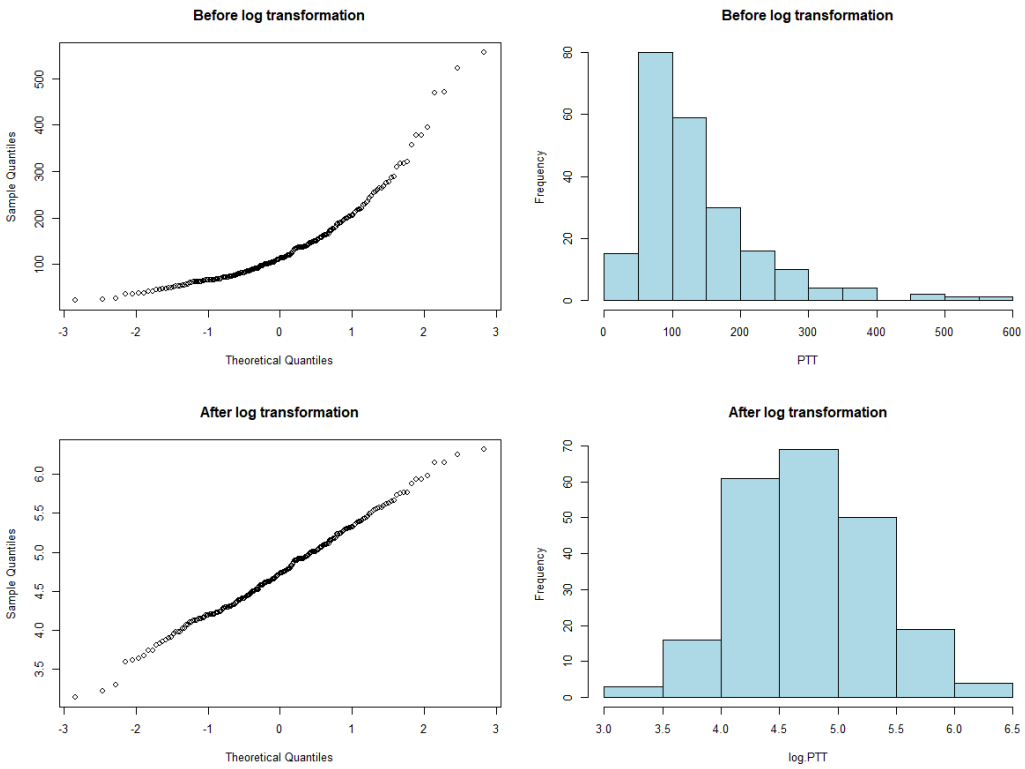

Fig 2.3 Distribution and Normality of PTT before and after the log transformation

2.5 Handling collinearity

Cases of collinearity between features were present in the dataset. In fact, in the case of the BFQ we have that traits score are just the sum of the facets scores and in the case of the positive and negative factor we have that their are just the weighted sum of the self-reported wayfindings inclinations features. Since collinearity is just redundant information and can case problems for some models we have decided to remove this features, ending this process with 22 features.

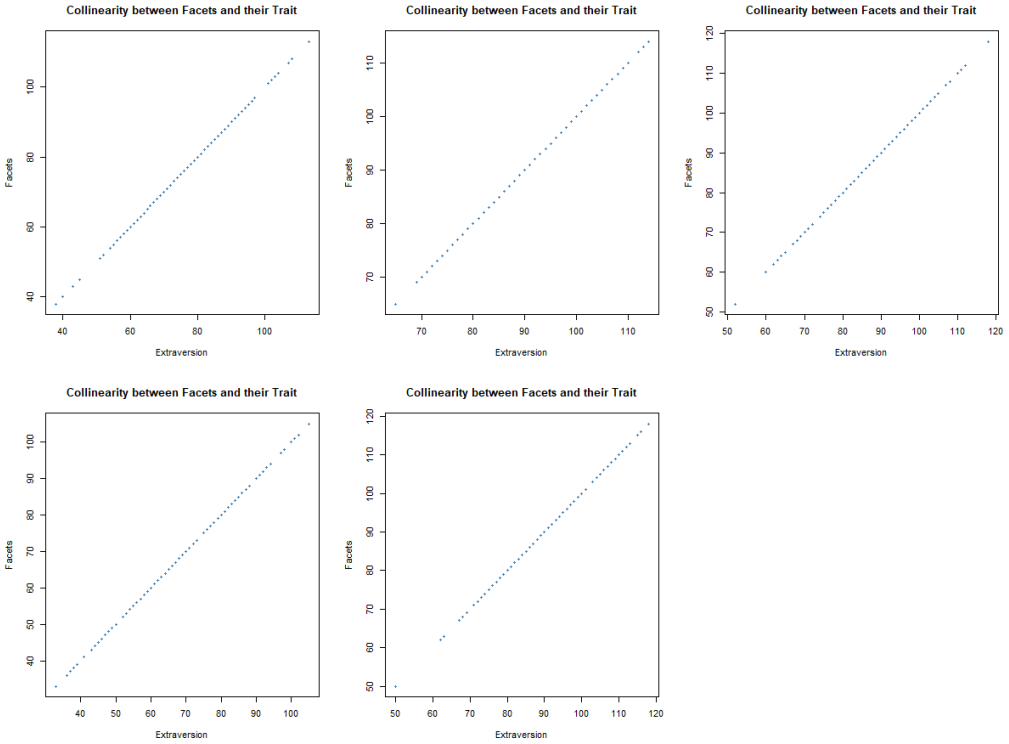

Fig 2.4 Collinearity between facets and trait

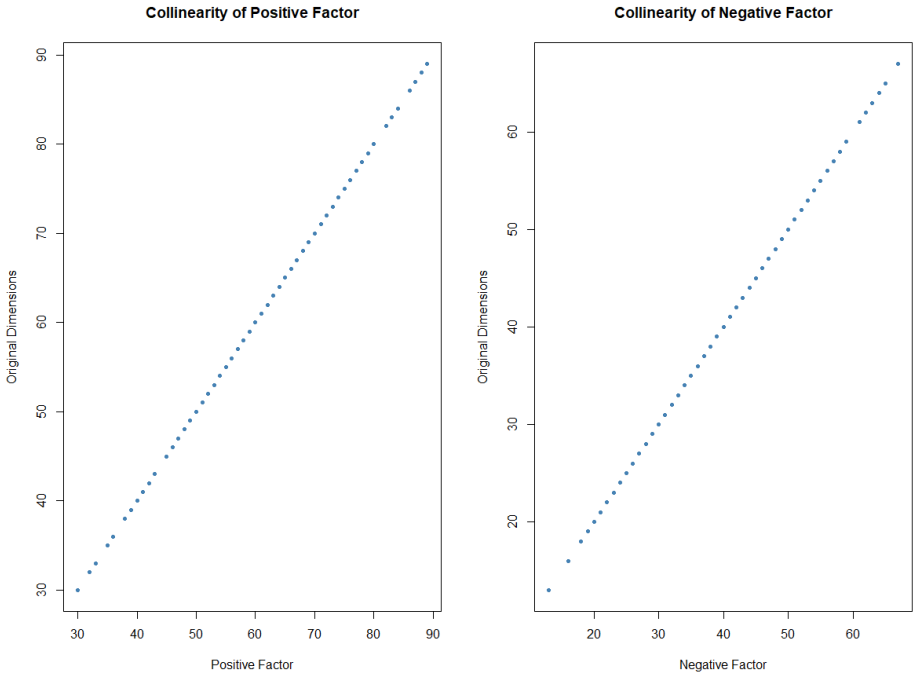

Fig 2.5 Collinearity between wayfindings and factors

2.6 Visualizing relationship between MRT and PTT and other features

Next step has been to visualize relationship between MRT and PTT and the others features with some plots with correlation coefficients and linear regression line for the numerical variables, while for the categorical ones a box-plot has been used.
Those first analysis showed sparse plots, low correlation coefficients and linear model with slope almost set to 0. This was useful in order to have a look at the strength of the relationship between our independent variables and our dependent ones and to estimate what we could expect from the classification models we used. The strongest relation, measured by correlation are the one with the *Cardinal Points* (PTT: -.29; MRT: .22) while the lowest correlations are with Experience Opening (PTT: .00; MRT: .02) and with Landmark and Route Mode (PTT: -.02; MRT: .00).

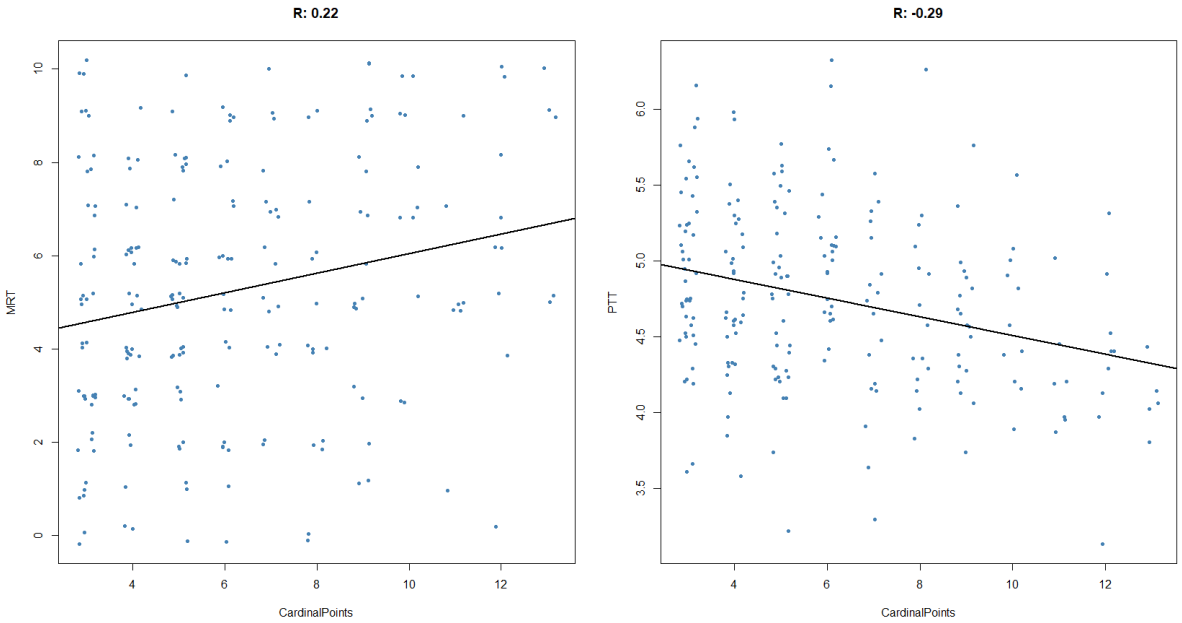

Fig 2.6 Plots of MRT and PTT with Cardinal Points

2.7 Transforming PTT and MRT in binary variables

Since MRT and PTT are just score which is goal is to measure the ability of the subjects, instead of trying to predict the score of PTT and MRT, we have tried to predict what PTT and MRT mean: high visuo-spatial abilities or low visuo-spatial abilities. For this reason PTT has been transformed to a binary variable where subject with lower score of the mean are set to 1 (high), since higher score of PTT represent a higher degree of errors and subject with higher score of the mean are set to 0. Vice-versa for MRT (higher score were set to 1, lower score were set to 0.
The previous analysis were repeated, but this time since we are dealing with categorical variables, box-plot were used for categorical-numerical relations (MRT-Cardinal Points) and bar-plot for categorical-categorical relations (MRT - Gender)

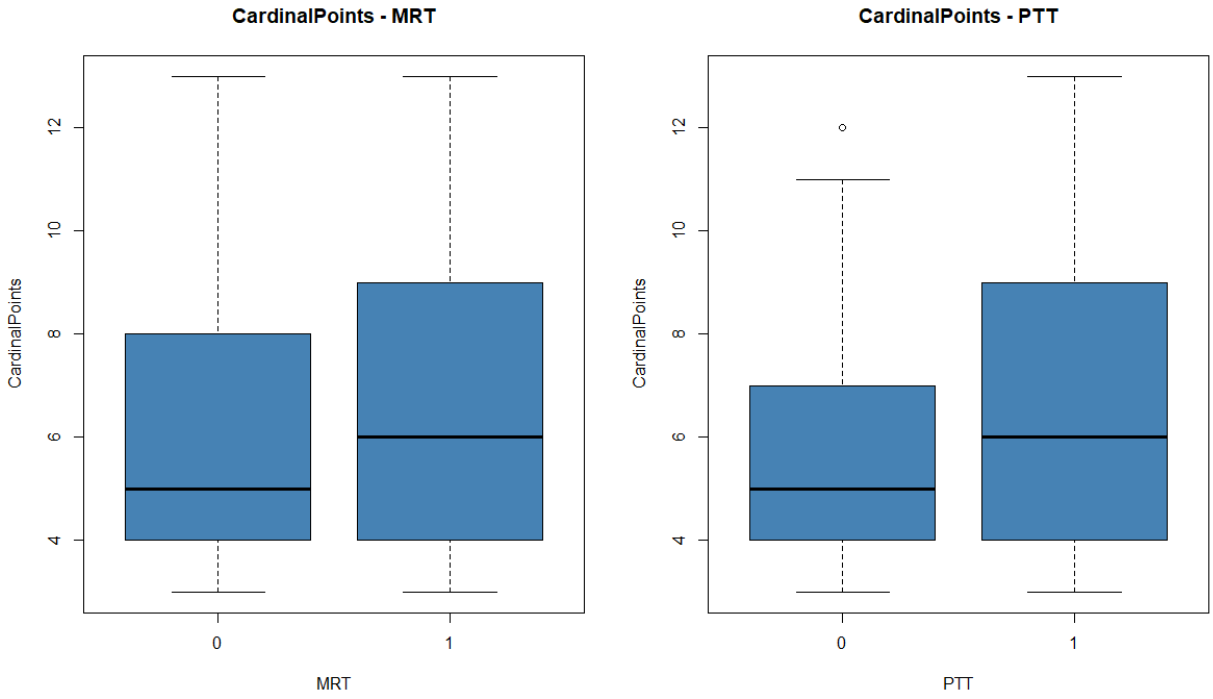

Fig 2.7 Relations of binary MRT and PTT with Cardinal Points.

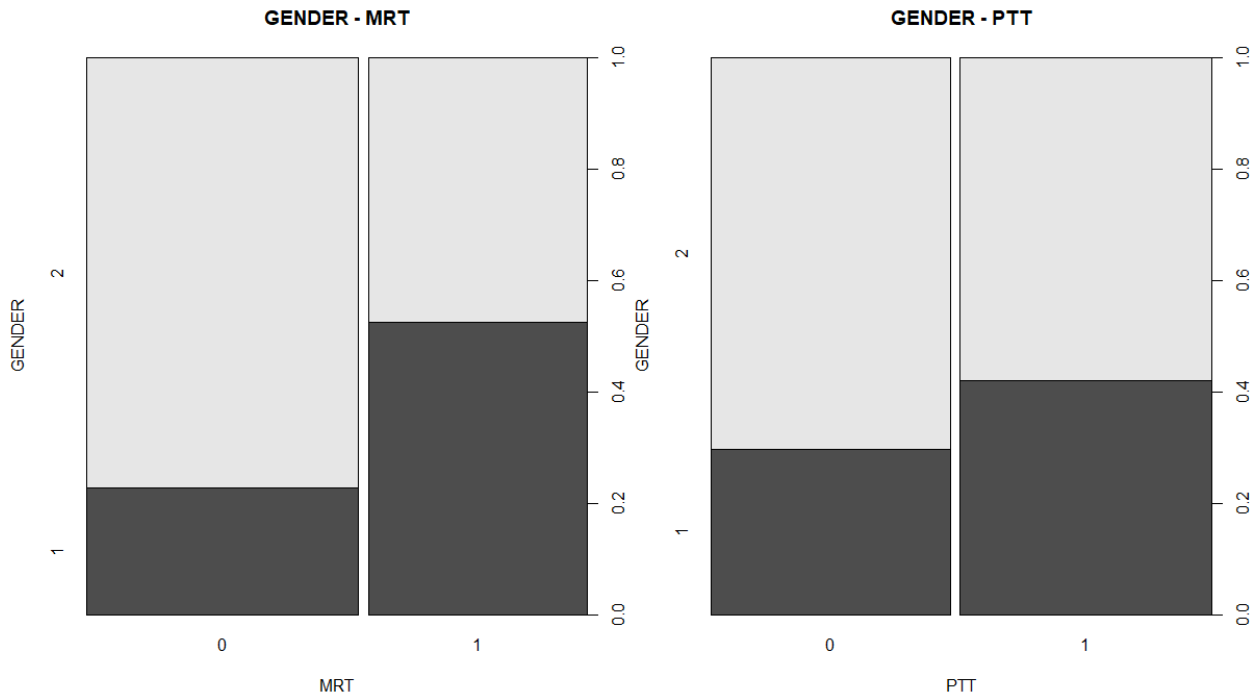

Fig 2.8 Relations of binary MRT and PTT with Gender. We can see how much Male (1) tends to have high MRT score (1).

Chapter 3

Models and features selection

3.1 Evaluating Measures

Different parameters were used to assess the goodness of the model.
1. *Accuracy*: the proportion of how many times the classifier guessed right, defined as the sum of *true positive (tp)* plus *true negative (tn)* over the sum of the number of *positive instances (P)* and the number of *negative instances (N)*.

$$Accuracy = \frac{tp + tn}{N + P}$$

2. *Sensibility and Specificity*: Those two are measured referred to the accuracy of the model for the single classes: *Sensibility*, also called *True Positive Rate (TPR)*, is the proportion of positive instances classified rightly, meanwhile *Specificity* refers to how many negative instances are right.

$$Sensibility = \frac{tp}{P} \tag{3.1}$$

$$Specificity = \frac{tn}{N} \tag{3.2}$$

3 Receiver operating characteristic Curve (ROC Curve): it's a graphical plot created by plotting the *true positive rate (TPR)* against the false *positive rate (FPR)* at various threshold settings.

$$FPR = \frac{fp}{N}$$

From the ROC Curve it has been computed the *Area Under the Curve (AUC)* which is literally the area under the ROC curve. An AUC greater than 0.5 indicates a classifier which works better than a random one, meanwhile an AUC of 1 indicates a perfect classifier.

4 To assess if the classifier works better than one with no information a *binomial test* has been performed where the number of success is the number of rightfully classified items, the total number of the sample is the total number of the instances. The null hypothesis is that the accuracy is greater than the proportion of the most numerous group. So that:

$$H0 : Accuracy > \frac{\max(P,N)}{N+P}$$

3.2 k-Nearest Neighbors: a non-parametric approach

3.2.1 The algorithm

The first model tried has been the k-Nearest Neighbors (k-NN). k-NN works this way: Given an instances of your data X(i), it checks the k nearest neighbours X(k) by computing the l2 norm of X(i) - X(k) and set Y(i) equals to the most numerous class of the set of all the k X(k).

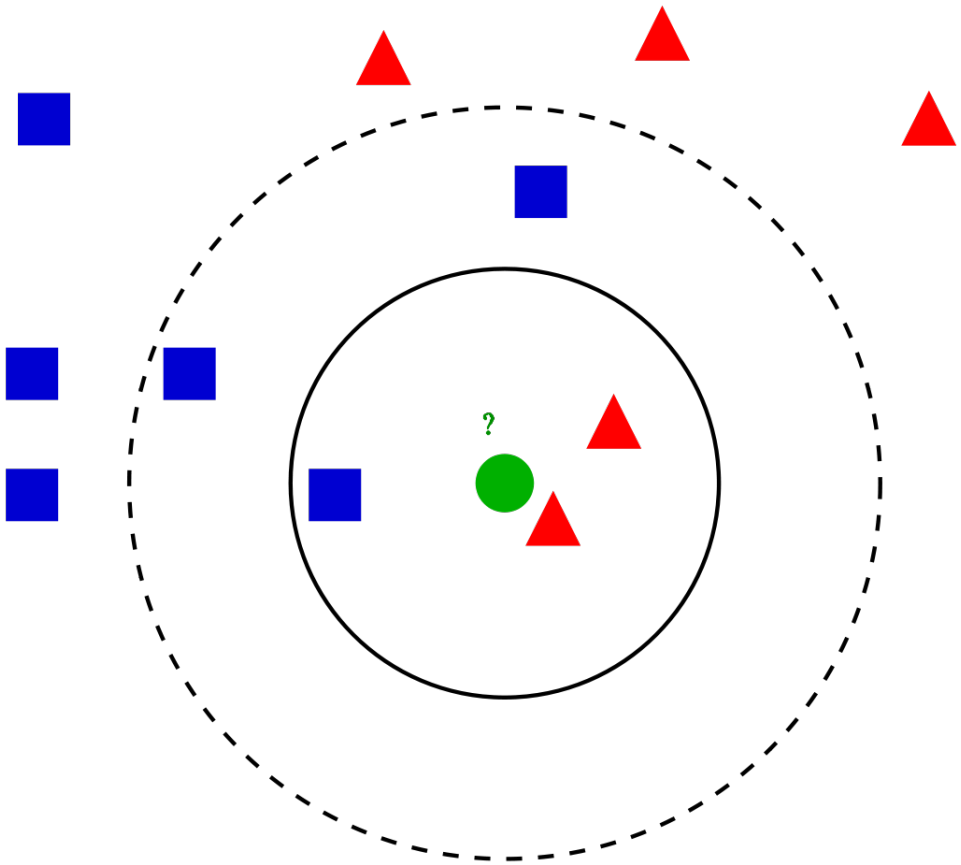

Fig 3.1 An example of k-NN with k equals 3 and 5

Other than being a very simple approach k-NN is suitable when your data doesn't belong to a space of high dimensions since in case of low dimensionality we don't have problems coming from the Gaussian Annulus Theorem, which says that in higher dimension the norm of our instances tends to have variance equals to 0.
It's important to note that k-NN needs standardized data, for this reason, since all the features seemed reasonably normal, the dataset has been standardized so that each features

would have been mean equals to 0 and variance equals to 1.

3.2.2 Leave-One-Out-Cross-Validation

Since the low number of instances (n=222) a Leave-One-Out-Cross-Validation (LOO-CV) has been used. LOO-CV is a special case of k-Fold-Cross-Validation when k is set to n, in this way LOO-CV uses n-1 instances as training data and 1 instance as validation data, this procedure is done for each instances. The computational burden is much higher than other approach but remain feasible when we deal with low number of instances.

3.2.3 Results

Different models were tested by changing the value of k from 1 to 25, meanwhile the features used were all the one available.
1. MRT: for MRT accuracy higher than No Information Rate Classifier (NIR-C) were reached only for k greater than 13.
The highest accuracy is 0.653 when k is set to 18 (Accuracy $>$ NIR, pvalue $<$ 0.003), meanwhile the AUC is 0.656 (C.I. = 0.587-0.724).

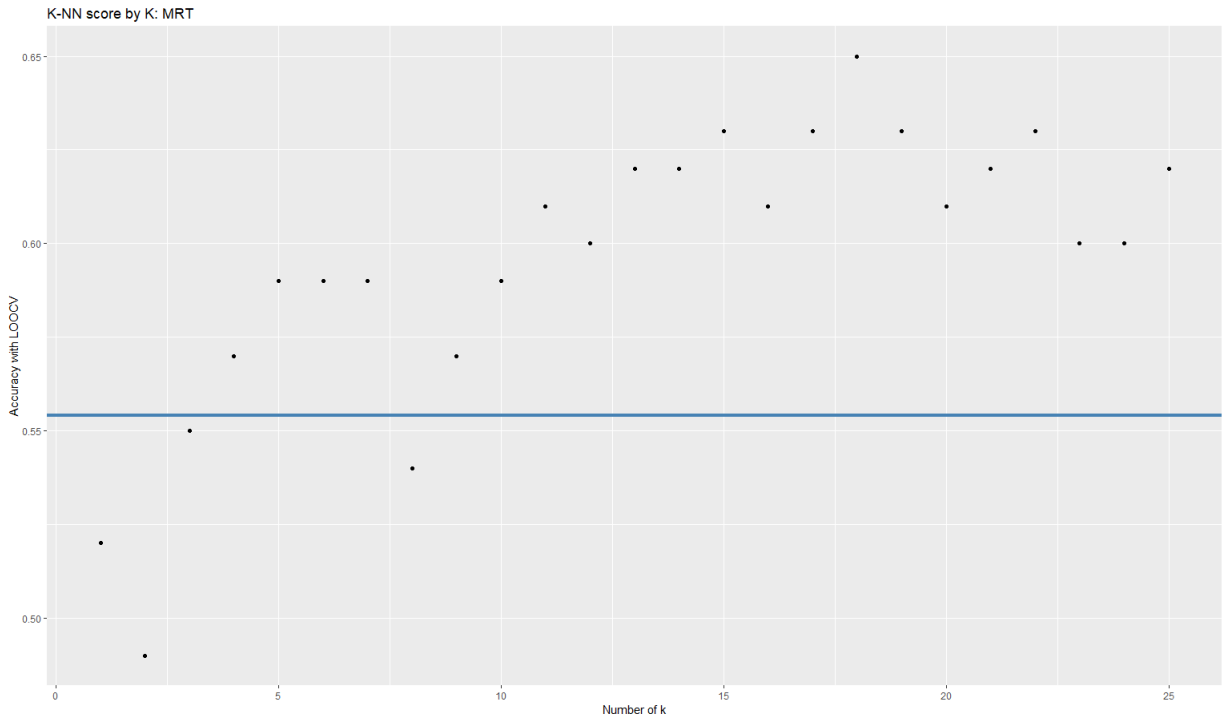

Fig 3.2 Accuracy of the kNN-MRT model by changing k

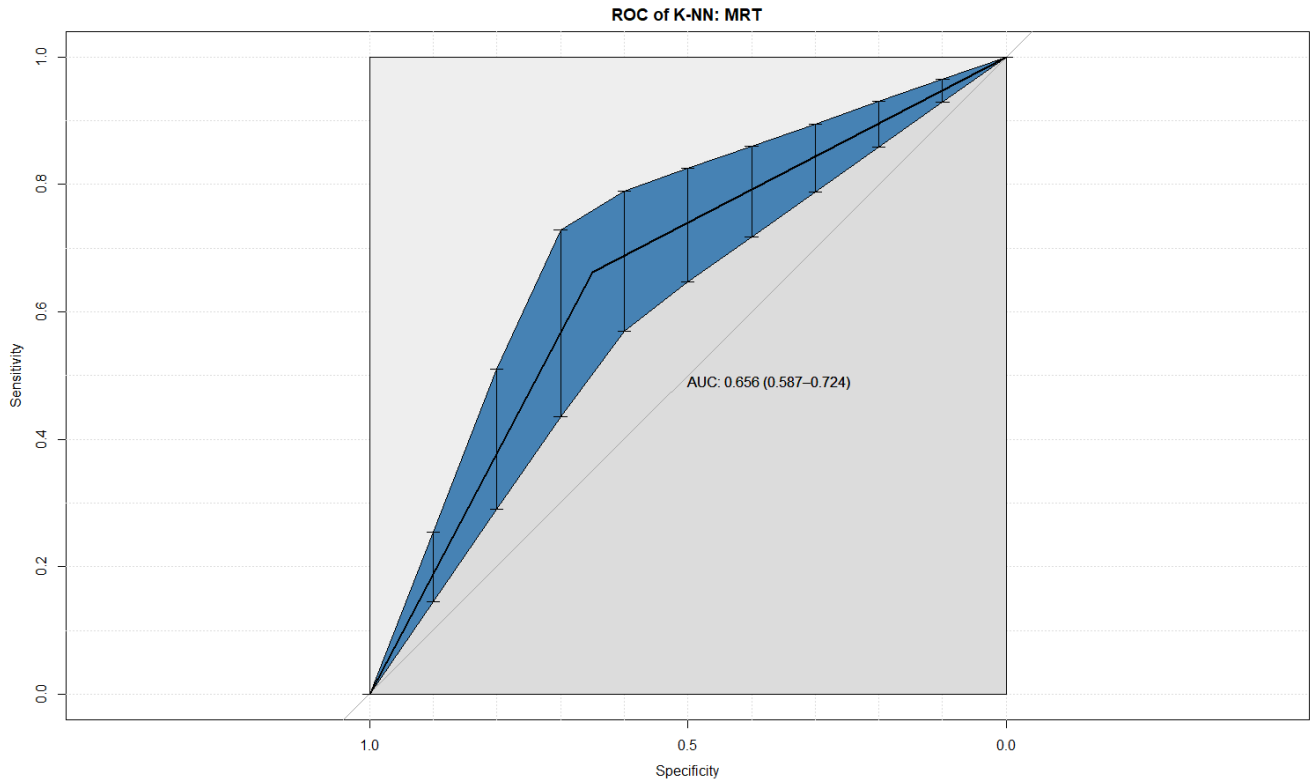

Fig 3.3 ROC Curve for the kNN-MRT model with k = 18

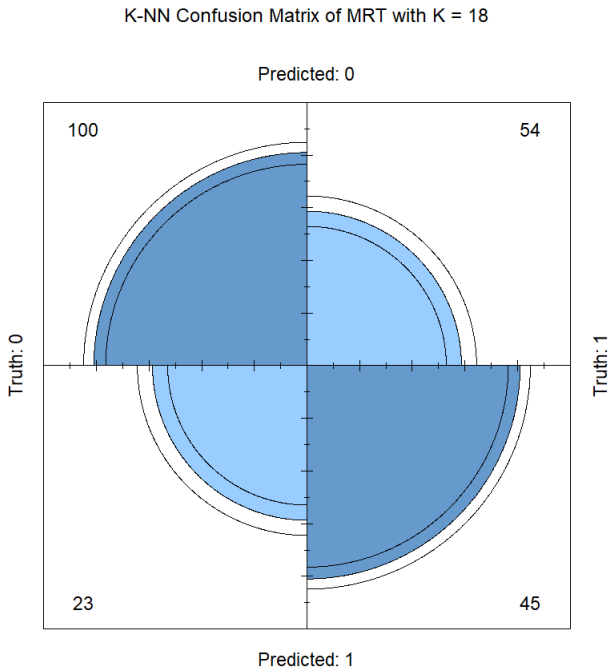

Fig 3.4 Confusion Matrix for the kNN-MRT model with k = 18

2. PTT: lower accuracy than for MRT has been found for PTT. The only models with

an Accuracy better than a one of a NIR-C are the ones with k set to 4, 18 and 20. When k is set to 4 the accuracy 0.581 (Accuracy > NIR, pvalue < 0.03) and an AUC of 0.584 (C.I. = 0.518-0.649).

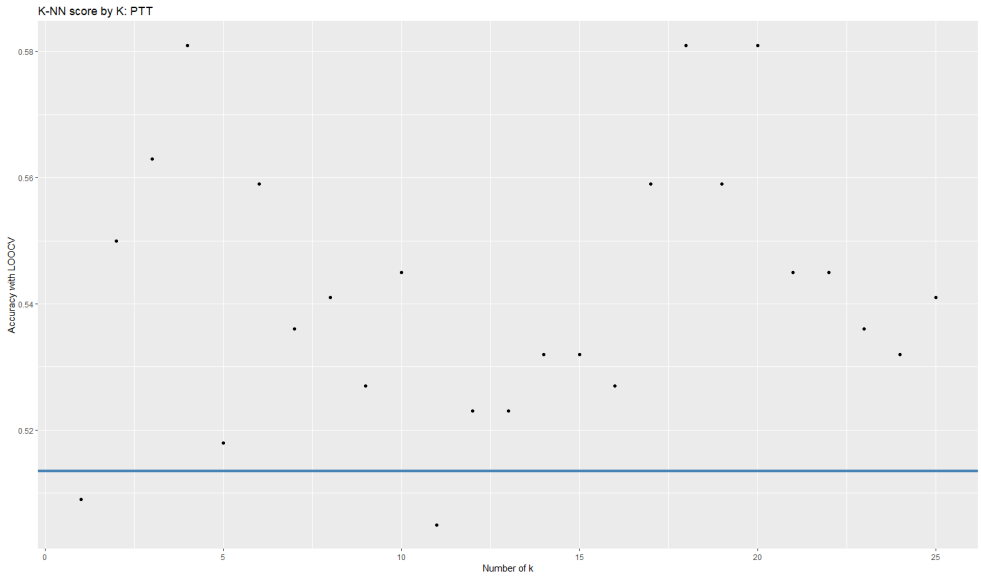

Fig 3.5 Accuracy of the kNN-PTT model by changing k

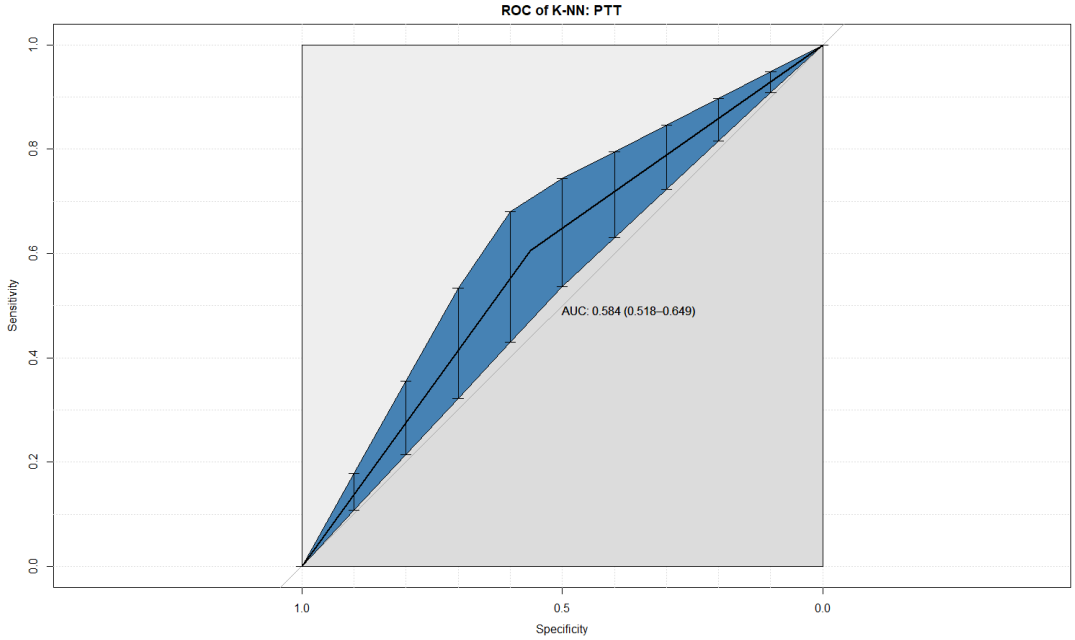

Fig 3.6 ROC Curve for the kNN-PTT model with k = 4

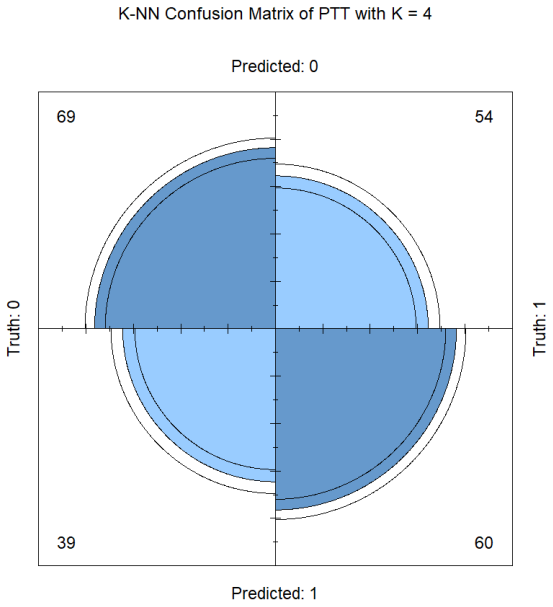

Fig 3.7 Confusion Matrix for the kNN-PTT model with k = 4

Results of a kNN model with all features haven't been high for both MRT and PTT, however MRT showed better results and also better consistency with different and high values of k.

3.3 Linear Discriminant Analysis

3.3.1 Description of the algorithm

Linear Discriminant Analysis (LDA) is a machine-learning techniques that can be applied in classification problems. The general idea is to use the training data to estimate trough maximum likelihood the best parameters for your multivariate Gaussian distribution (3.1) with the assumption that they have equals covariance matrix Σ.

PDF of a multivariate Gaussian distribution

$$f(y_i|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} exp(-\frac{1}{2}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)) \tag{3.3}$$

Then we can compute the posterior probabilities of one instances to belonging to one of our g distributions (3.2) . We assign that instance to the class with the highest posterior

probability. Note that prior probabilities are taken into account, in the case of a binary classifier the prior probabilities are P/(N+P) and N/(P+N).

Posterior probabilities of x belonging to class j

$$P(G_j|x) = \frac{\frac{\pi_j}{\sqrt{(2\pi)^n|\Sigma|}} exp(-\frac{1}{2}(x-\mu_j)'\Sigma^{-1}(x-\mu_j))}{\sum_{i=1}^{g} \frac{\pi_i}{\sqrt{(2\pi)^n|\Sigma|}} exp(-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i))} \tag{3.4}$$

The problem can be simplified even further taking the logarithm of the posterior probabilities so that we assign x to the class with the highest discriminant score (3.3).

Discriminant Score function for LDA

$$\delta_j(x) = x^T\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j + \log\pi_j \tag{3.5}$$

3.3.2 Results for full features LDA model

Accuracy and AUC were measured with a LOO-CV procedure. for both MRT and PTT 1. MRT: for MRT the accuracy of LDA was 0.626 showing to perform just above of a random classifier (Accuracy $>$ NIR, pvalue $<$ 0.03), moreover the classifier showed a tendency to classify towards the negative class (low MRT score/low spatial abilities). The AUC score was a little better than the AUC of a random classifier is 0.620 (C.I. = 0.554-0.686).
See figures to see confusion matrix and ROC curve.

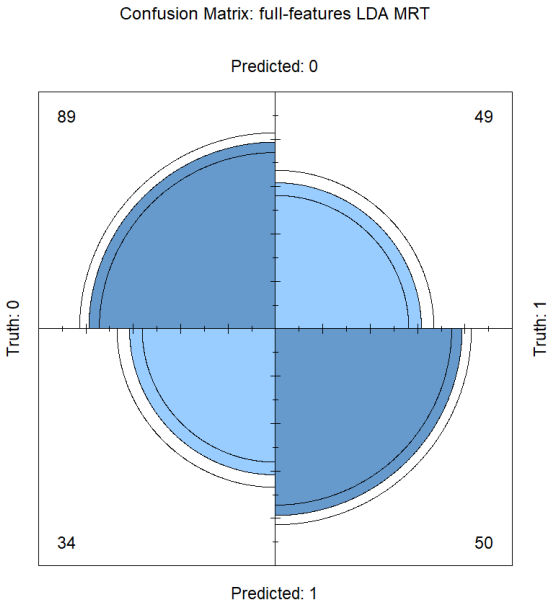

Fig 3.8 Confusion matrix of LDA-MRT full model

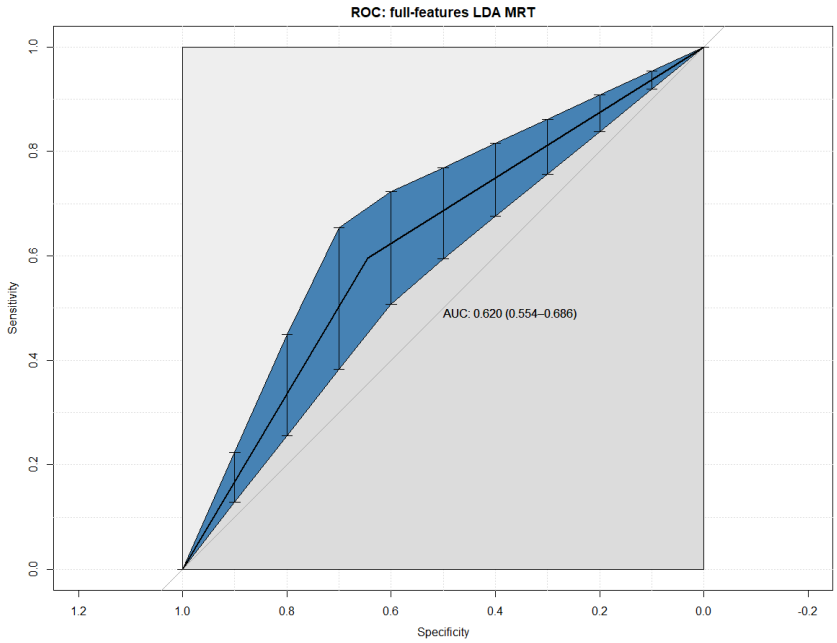

Fig 3.9 ROC curve and AUC of LDA-MRT full model

1. PTT: the model for PTT was not be able to perform better then a random classifier. Accuracy was 0.518 (Accuracy > NIR, pvalue > 0.47). The AUC of the model was 0.518 (C.I. = 0.452-0.584).

3.3.3 Features Selection

Since the bad results with a full features model, we decided to perform a greedy features selection based on the relative differences (3.4) of the two posterior means components of the multivariate Gaussian distributions.
We started with a model with only the most separated feature, then we kept on with a two-features model (the previous one and the second most separated feature. We did this procedure until we reached the full models. Note that the accuracy of every model was calculated with a LOO-CV procedure.

Separation score for features i

$$R_i = \frac{|\mu_{i,1} - \mu_{i,2}|}{\mu_{i,1} + \mu_{i,2}} \tag{3.6}$$

1. MRT: for MRT higher accuracy than a full features model has been reached with few features. In particular, the nest accuracy has been reached with six and seven features (Accuracy = 0.671).

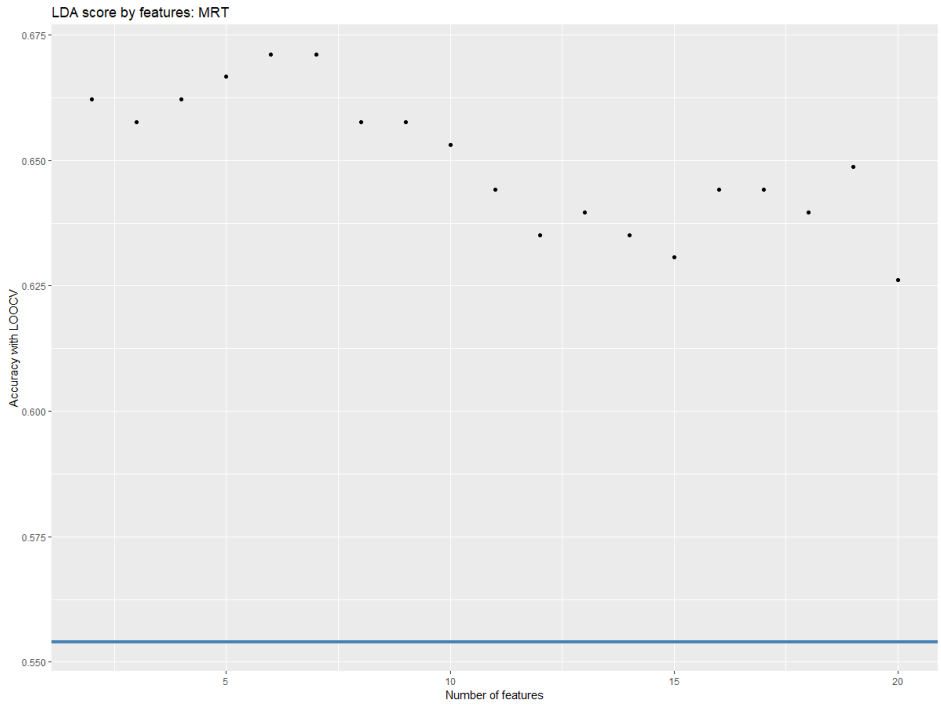

Fig 3.10 Accuracy for MRT-LDA by number of features

The best features are: Gender, Emotion Control, Pulse Control, QAS, Cardinal Points and Sense of Direction. The model was able to have an reach an high specificity (*specificity = 0.789*) but a low sensitivity (*sensitivity = 0.526*).

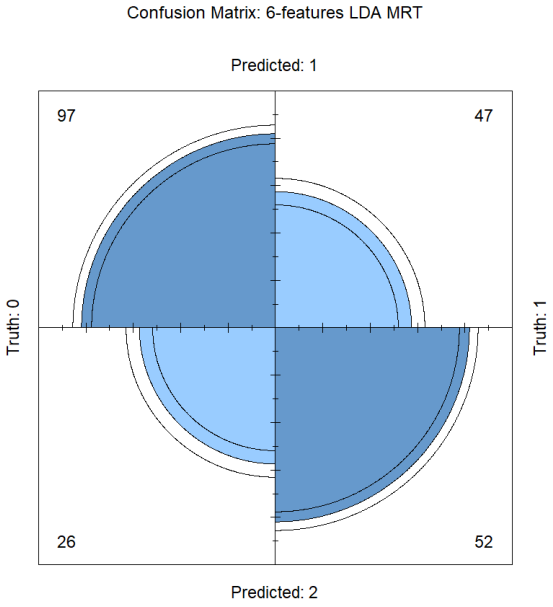

Fig 3.11 Confusion Matrix of MRT-LDA model with six features

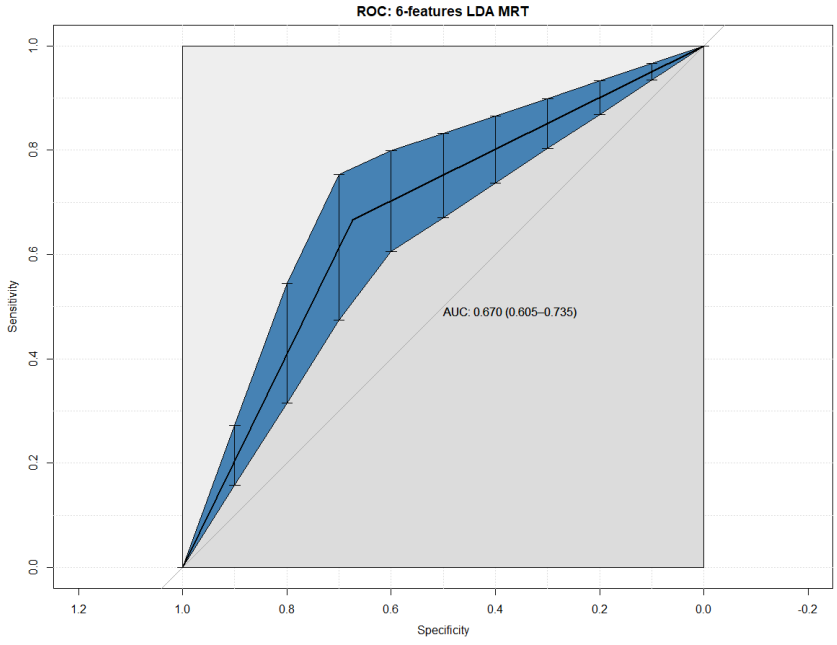

Fig 3.12 ROC curve and AUC of MRT-LDA model with six features

2- PTT: for PTT the same effect of MRT was found: less features meant higher accuracy.

In this case the best PTT model have a 0.595 Accuracy with four features: Gender, Map Use, Verbal Indication and Cardinal Points.

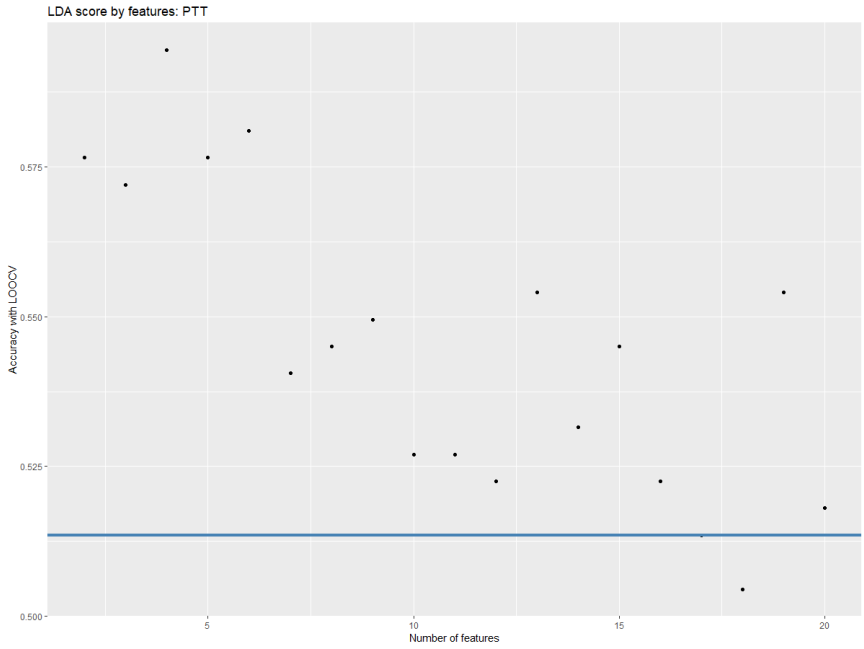

Fig 3.13 Accuracy for PTT-LDA model with four features

The model showed to have equal almost *specificity* and *sensitivity* (respectively: 0.611 and 0.579).

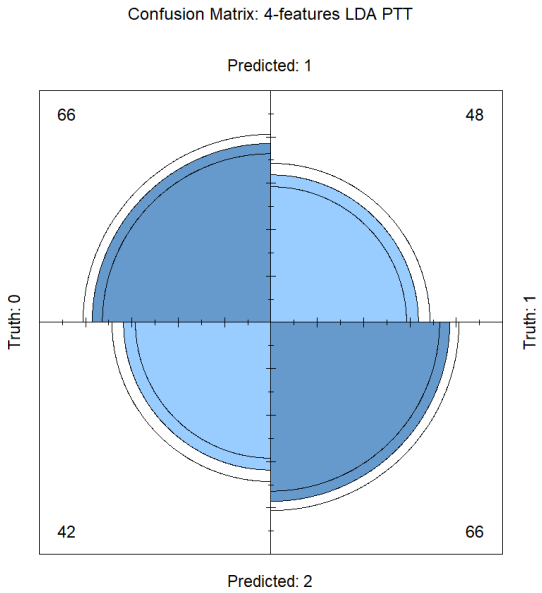

Fig 3.14 ROC curve and AUC of PTT-LDA model with four features

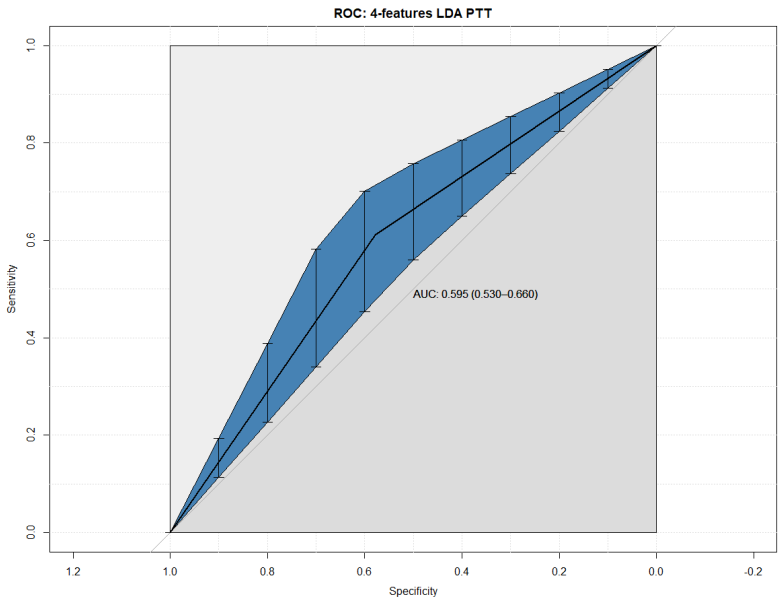

Fig 3.15 ROC curve and AUC of PTT-LDA model with four features

3.3.4 LDA final results

LDA showed to perform really poorly with a full features model (Accuracy MRT: 0.626, Accuracy PTT: 0.518). But with a features selection based of the separation score of the means of the features we managed to reach better Accuracy (MRT: 0.71, PTT: 0.595) than a full features model or a k-NN model (significant differences were not checked). It's important to note that for both MRT and PTT Gender and Cardinal Points were included as features for the models. In the MRT model were included personality traits regarding the emotional stability (Emotion Control and Pulse Control) together with some self-assessment wayfindings inclinations (QAS and Sense of Direction), while in the PTT model the others features came from the NAQ (Map Use and Verbal Indication).

3.4 Quadratic Discriminant Analysis

3.4.1 Description of the algorithm

Quadratic Discriminant Analysis (QDA) is based on the same principle of LDA, but with one main difference: if in LDA the variance is the same for all g multivariate Gaussian distribution, in QDA the variance is not set to be equal for all the distributions. This means that we need to estimate for each one of our g distributions both means and variance, implying a different discriminant score (3.5). The positive effect of computing

the variances is that our boundaries can be non-linear, the negative effect is the risk of over-fitting is higher than using LDA.

Discriminant Score function for QDA

$$\delta_j(x) = x^T\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j - \frac{1}{2}x^T\Sigma^{-1}x + \log\pi_j \tag{3.7}$$

3.4.2 An over-fitting problem

As previously said QDA is a much powerful algorithm then LDA and this can cause major over-fitting problems. In fact, we initially tested QDA for both MRT and PTT without using a cross-validation technique, looking up only for the difference between training accuracy and LOO-CV accuracy: the result showed major differences between training and LOO-CV accuracy, the accuracy in training set for MRT and PTT were respectively 0.829 and 0.869 while the accuracy with LOO-CV were 0.527 and 0.523 and these classifiers were not better than NIR classifiers (p.value MRT $>$ 0.81, p.value PTT $>$ .42).

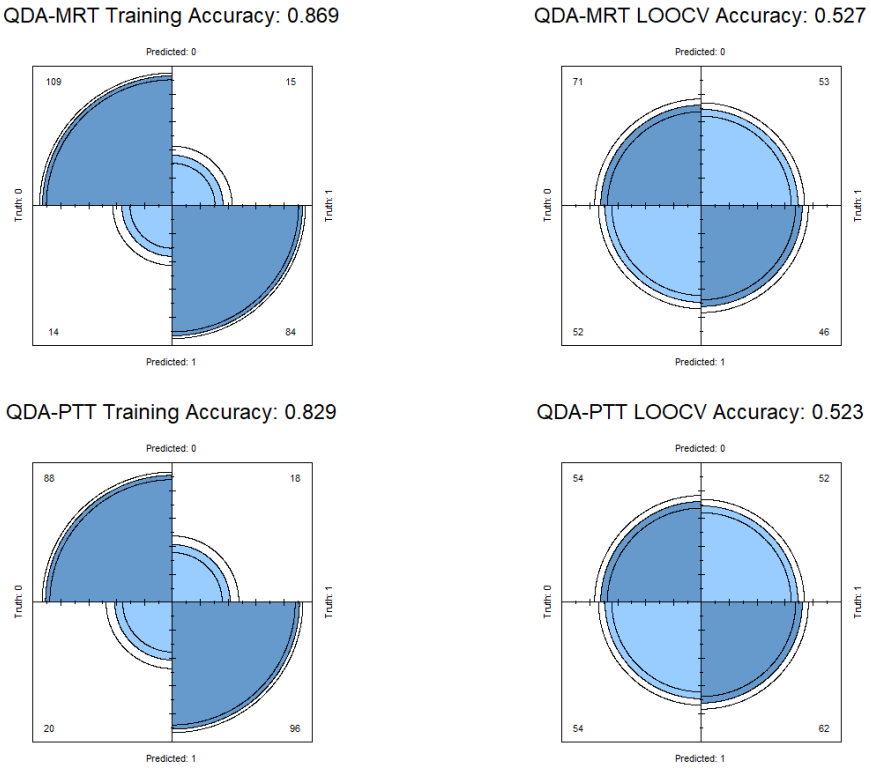

Fig 3.16 Comparison of training vs validation accuracy in a QDA

3.4.3 Features Selection

To try to improve the low accuracy of the QDA model we tried to perform a feature selection in order to remove features which were adding only noise and not information. The procedure of feature selection is the same used for LDA: try to select the most separated features.

1. MRT: in this case the best model in terms of accuracy was one with only two features: Gender and Cardinal Points. It reached an accuracy of 0.653, with high specificity (0.756) but low sensitivity (0.526). The AUC was 0.649 (C.I. = 0.584-0.715).

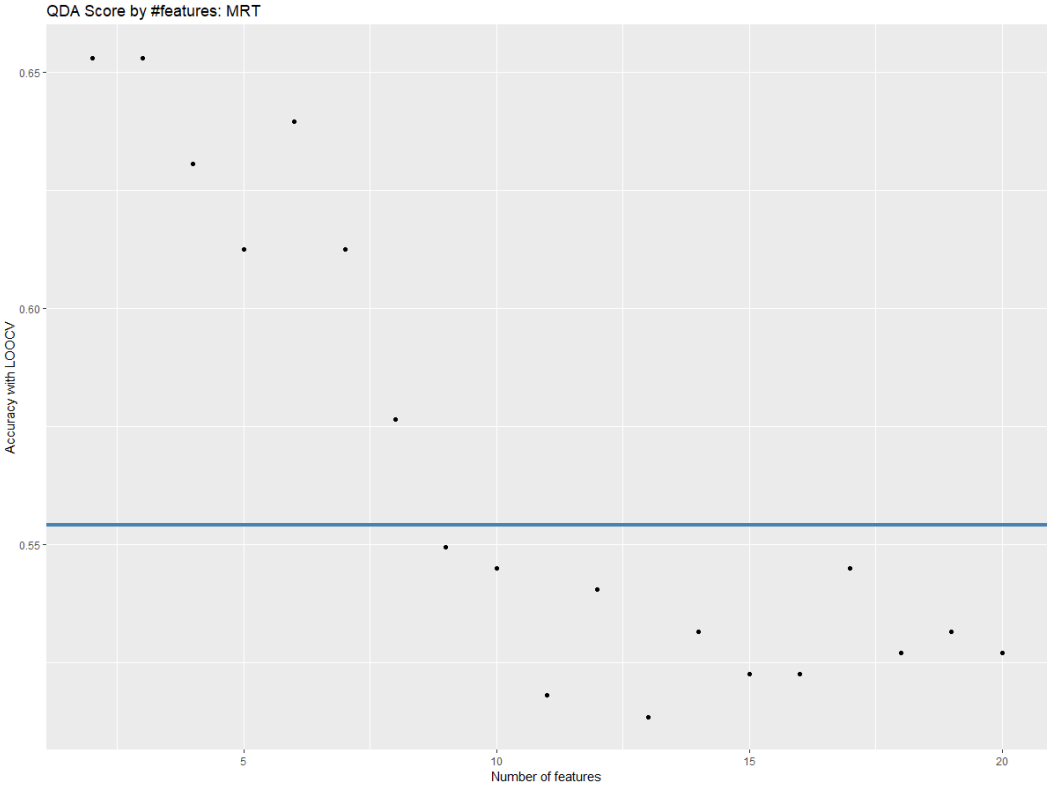

Fig 3.17 Accuracy by numbers of features in MRT-QDA

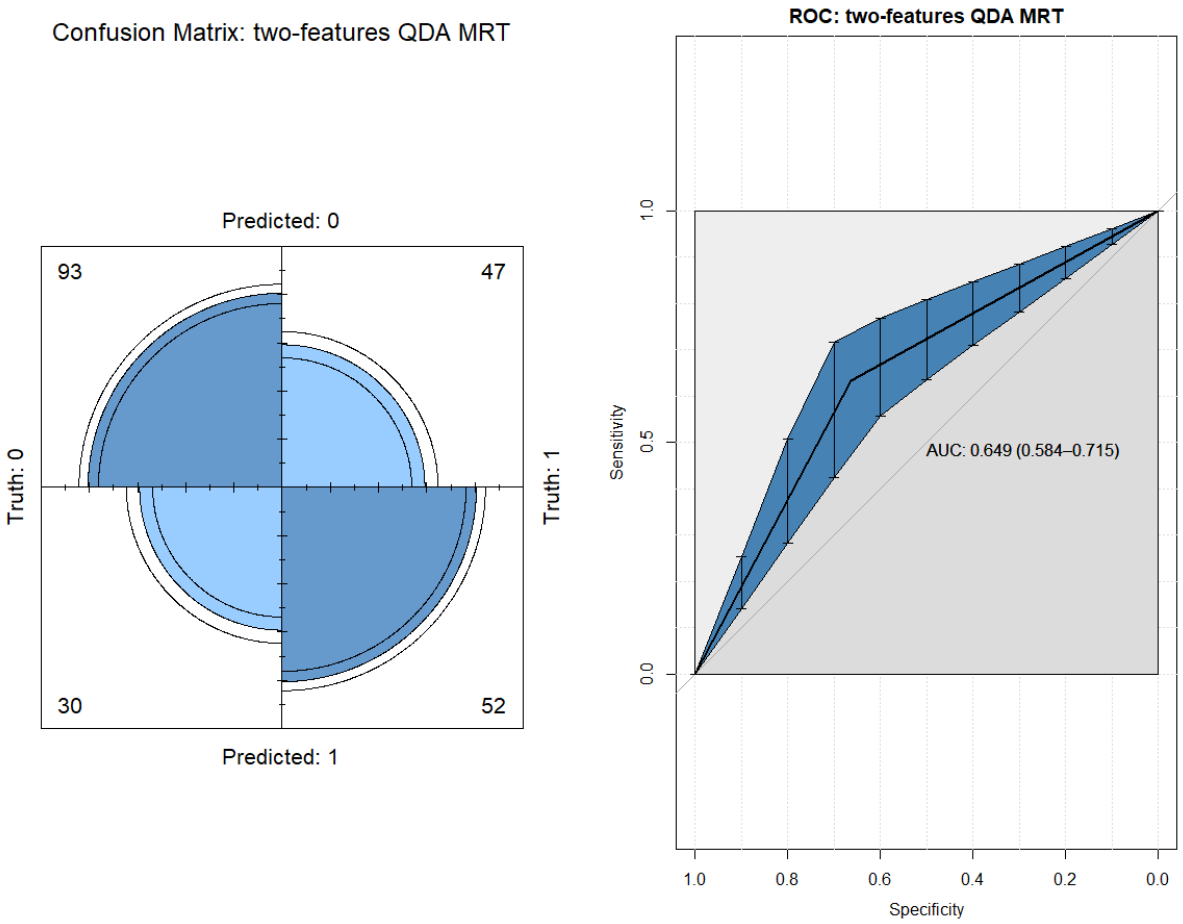

Fig 3.18 Confusion Matrix and ROC curve for MRT-QDA

2. PTT: as in previous model PTT showed lower accuracy than MRT. The best accuracy was reached with four variables, the same of the LDA model: Gender, Map Use, Verbal Indication, Cardinal Points. Switching from all the features to four increased the accuracy form 0.523 to 0.559.
Still, this accuracy is not significantly higher than an accuracy of a NIR classifier (Accuracy $>$ NIR: p.value $>$ 0.10). The specificity of the model was 0.593 and the sensitivity was 0.526, with an AUC of 0.560 (C.I. = 0.494-0.625).

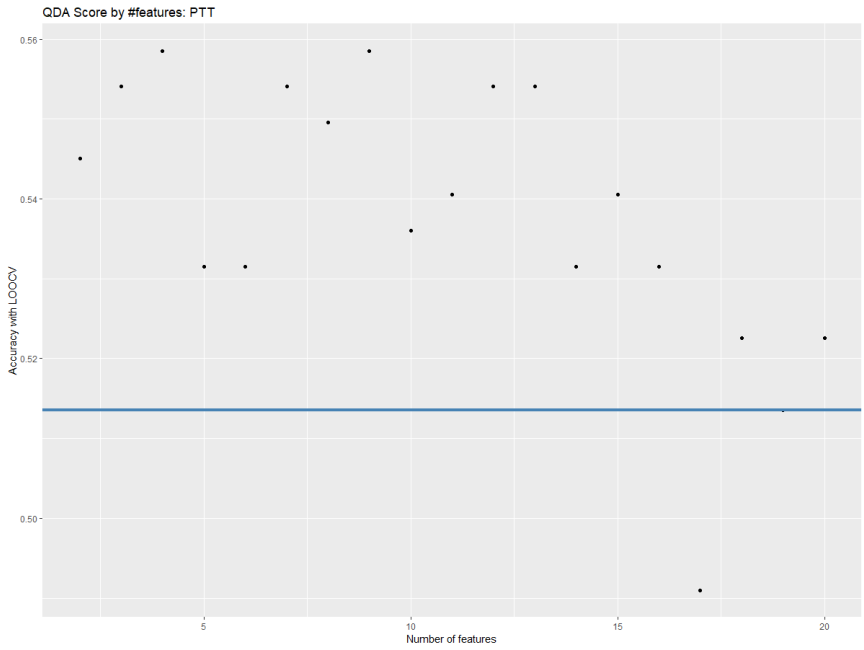

Fig 3.19 Accuracy by numbers of features in PTT-QDA

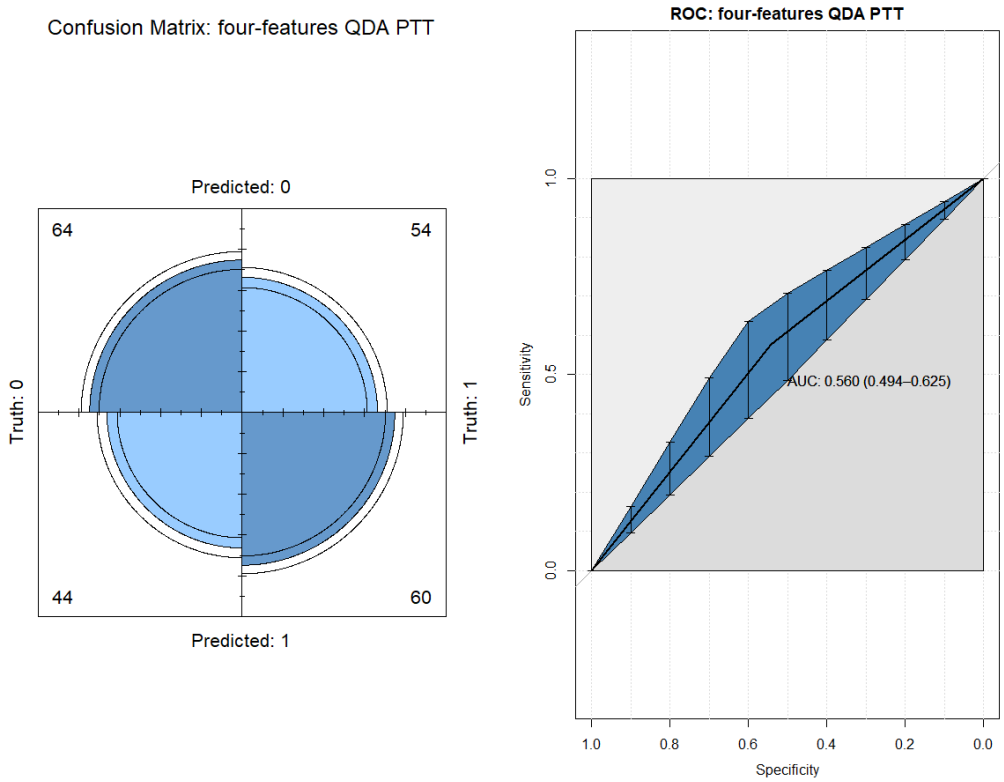

Fig 3.20 Confusion Matrix and ROC curve for PTT-QDA

3.5 Generalized Linear Model: Logistic Regression

When the response variable Y is binary, like in our case, a standard linear model is not the most suitable model of regression. In fact in a linear model we have that E[Y] = βX where Y could easily fall out from our possible value [0, 1]. To fix this problem we use a link function which relates the mean of our distribution, in our case a Bernoulli Random Variable such that E[Y] = π_i.
In this case our link function is:

$$logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i}) \tag{3.8}$$

So that our model is specified in this way:

$$log(\frac{\pi_i}{1-\pi_i}) = \beta X \tag{3.9}$$

And by isolating π_i we obtain:

$$\pi_i = \frac{exp(\beta X)}{1+exp(\beta X)} \tag{3.10}$$

We can observe that now our regression model resemble a binary classification model, where the predicted value (3.8) can be interpreted as the probability to belong to the positive class, so:

ds

$$P(Y_i \in High) = P(Y_i = 1) = \frac{exp(\beta X)}{1+exp(\beta X)} \tag{3.11}$$

$$P(Y_i \in Low) = P(Y_i = 0) = \frac{1}{1+exp(\beta X)} \tag{3.12}$$

The model can be fitted by maximizing the log-likelihood:

$$l(\beta; y) = log(\prod_{i=1}^{n}(\pi_i^{y_i} + (1-\pi_i)^{(1-y_i)})) = \sum_{i=1}^{n}(y_i log(\pi_i) + (1-y_i) log(1-\pi_i)) \tag{3.13}$$

Since the distribution of $\hat{\beta}$ are approximately normal the distribution:

$$Z_j = \frac{\hat{\beta}_j - \beta}{\hat{SE}(\hat{\beta}_j)} \tag{3.14}$$

is approximately standard normal. This can hence be used for testing the hypotheses that individual coefficients are zero and to construct confidence intervals for individual parameters.

3.5.1 Full-Features models

We started by applying a GLM with all the features available, for both our response variables: MRT and PTT.

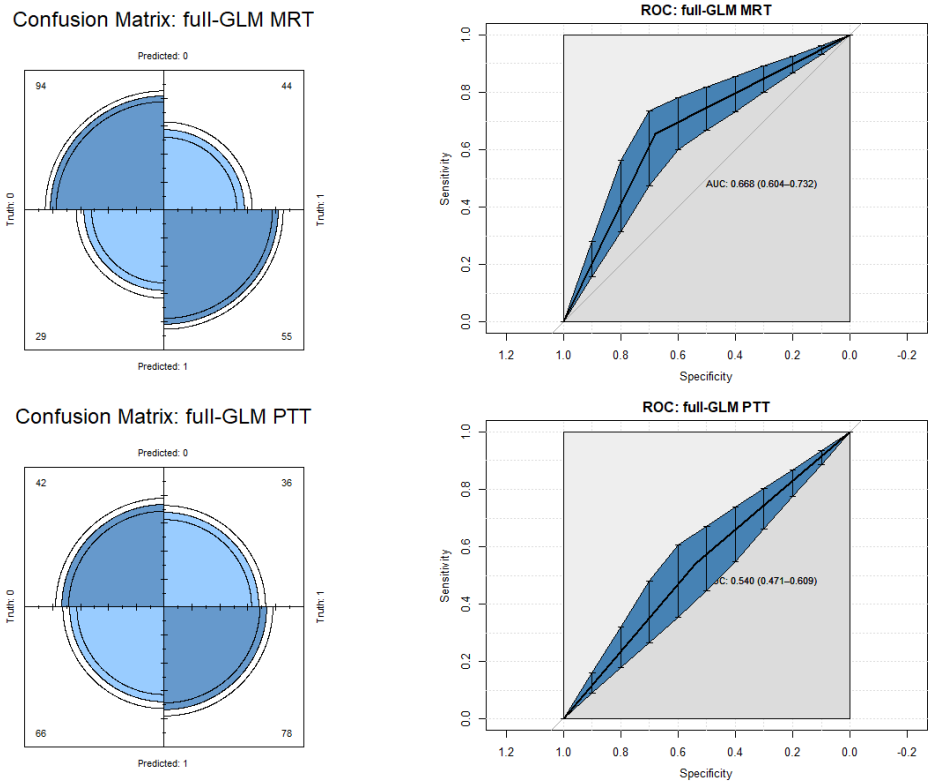

Fig 3.21 Confusion Matrix and ROC curve for MRT and PTT GLM's

1. MRT: The accuracy reached for MRT-GLM is 0.671, with a sensitivity of 0.556 and a specificity of 0.764. The AUC was 0.668 (C.I. = 0.604-0.732). However this effect seems to be provided by only the variable Gender (Z: -3.29; p.value < 0.001). The others features seemed to add only noise to the model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.5373	2.6111	-0.59	0.556
GENDER	**-1.2075**	**0.3666**	**-3.29**	**0.001**
X1.DYNAMISM	-0.0103	0.0328	-0.31	0.754
X1.DOMINANCE	0.0130	0.0272	0.48	0.633
X2.COOPERATIVITY	-0.0362	0.0409	-0.89	0.376
X2.CORDIALITY	0.0449	0.0329	1.37	0.172
X3.SCRUPOLOSITY	0.0005	0.0256	0.02	0.985
X3.PERSEVERANCE	-0.0030	0.0247	-0.12	0.902
X4.EMOTION.CONTROL	0.0254	0.0245	1.04	0.299
X4.PULSE.CONTROL	0.0196	0.0270	0.72	0.469
X5.CULTURE.OPENING	0.0349	0.0259	1.35	0.177
X5.EXPERIENCE.OPENING	-0.0268	0.0288	-0.93	0.353
Map.Use	-0.0458	0.1100	-0.42	0.677
GPS.Use	0.0108	0.1312	0.08	0.934
Verbal.Indication	0.0155	0.1273	0.12	0.903
QAS	-0.0159	0.0326	-0.49	0.626
QACOexploration	0.0460	0.0560	0.82	0.412
QACOknown	0.0069	0.0451	0.15	0.878
CardinalPoints	0.0047	0.0662	0.07	0.944
SenseOfDirection	-0.0045	0.0468	-0.10	0.924
LandmarkRouteMode	-0.0374	0.0722	-0.52	0.605

Table 3.1 Summary of MRT-GLM full features model.

2. PTT The accuracy reached for PTT-GLM is 0.540, also in this case this accuracy was the best so far for PTT. The sensitivity is 0.684, while the specificity is much lower (specificity = 0.389). The AUC was 0.540 (C.I. = 0.471-0.609). In this case this accuracy seemed to be provided by two variables: Experience Opening (Z: -2.40; p.value < 0.017) and Cardinal Points (Z: 2.55; p.value < 0.011).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.3721	2.5973	0.91	0.361
GENDER2	-0.4255	0.3658	-1.16	0.245
X1.DYNAMISM	0.0439	0.0315	1.39	0.164
X1.DOMINANCE	-0.0289	0.0265	-1.09	0.276
X2.COOPERATIVITY	-0.0431	0.0386	-1.11	0.265
X2.CORDIALITY	0.0420	0.0320	1.31	0.189
X3.SCRUPOLOSITY	0.0011	0.0248	0.04	0.966
X3.PERSEVERANCE	-0.0259	0.0246	-1.05	0.292
X4.EMOTION.CONTROL	0.0259	0.0238	1.09	0.276
X4.PULSE.CONTROL	-0.0152	0.0265	-0.57	0.567
X5.CULTURE.OPENING	0.0449	0.0251	1.79	0.074
X5.EXPERIENCE.OPENING	**-0.0685**	**0.0285**	**-2.40**	**0.017**
Map.Use	0.0724	0.1097	0.66	0.509
GPS.Use	-0.0449	0.1299	-0.35	0.730
Verbal.Indication	-0.1125	0.1248	-0.90	0.367
QAS	-0.0057	0.0318	-0.18	0.857
QACOexploration	0.0464	0.0551	0.84	0.400
QACOknown	-0.0140	0.0439	-0.32	0.750
CardinalPoints	**0.1689**	**0.0662**	**2.55**	**0.011**
SenseOfDirection	-0.0557	0.0470	-1.19	0.236
LandmarkRouteMode	-0.0546	0.0707	-0.77	0.440

Table 3.2 Summary of PTT-GLM full features model.

3.5.2 Features Selection

3.5.2.1 Procedure

In order to select the best features a step-forward features selection approach has been used. This approach uses as evaluating measure not the Accuracy but the *Akaike information criterion* (AIC). Since models like GLM works as "more features, better results" even if the added features are just random noise, a measure to evaluate the model based both the goodness of fit and the number of parameters is preferred.
The AIC accomplish this goal by being computed this way:

$$\text{AIC} = 2k - 2\ln(\hat{L}) \tag{3.15}$$

where:
k = *number of features*
$\hat{L}$ = *maximum value of the likelihood function of the model*

Now, that we explained what the AIC is and how it is computed, we are going to explain how our features selection algorithm works.

AIC-based Step Forward Features Selection :
1. Set *features*, $k = 0$, 0 and $F :=$ set of all features
2. Set n $= |F|$ 3. For every features $f_i \in F$ we compute:

$$\text{AIC}_i(\textit{response variable} \sim \textit{features} + f_i)$$

4. We now set *features* = *features* + f_j s.t:

$$\text{AIC}_j(\textit{response variable} \sim \textit{features} + f_j) \leq$$
$$\text{AIC}_i(\textit{response variable} \sim \textit{features} + f_i)\ \forall i$$

5. $k = k + 1$
6. We set $F = F \backslash f_j$
7. If k $>$ n:

Return *features*

else:

Go to step 3

3.5.2.2 Results

1. MRT: for MRT the lowest AIC's were with few features model. The lowest AIC was 284.44 , when only Gender and Emotion Control were used as features. The highest AIC was when all the features were used (AIC: 312.44).

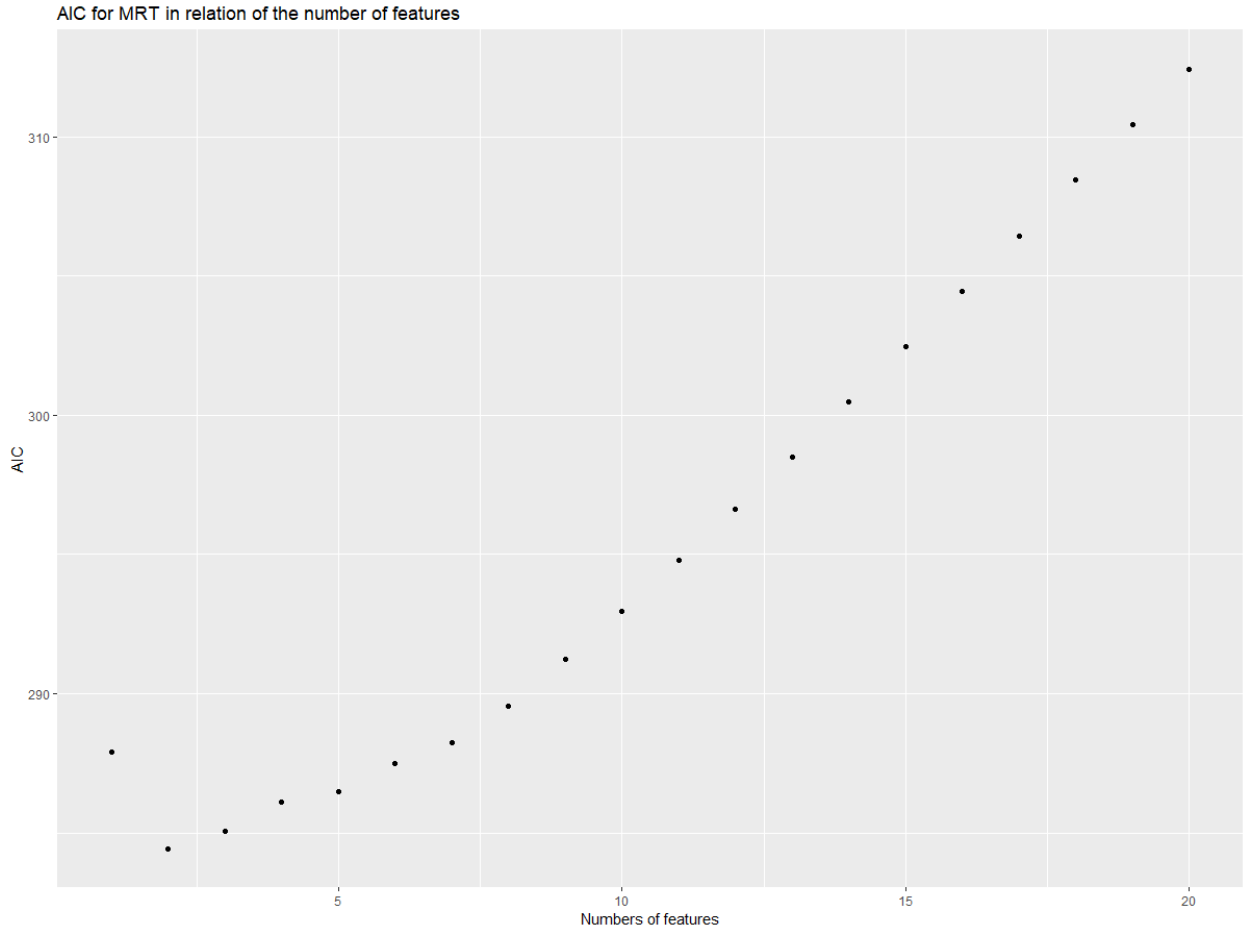

Fig 3.22 AIC's by numbers of features in MRT-GLM

Starting from this AIC's information we tested a Logistic Regressor Model with Gender and Emotion Control as predictors and MRT as response variable.
The accuracy of a model using LOO-CV as validation procedure showed a growth of accuracy compared to the full-features model. The accuracy was 0.676, with a sensibility of 0.535, a specificity of 0.789 and an AUC of 0.674 (C.I. = 0.609 - 0.738).

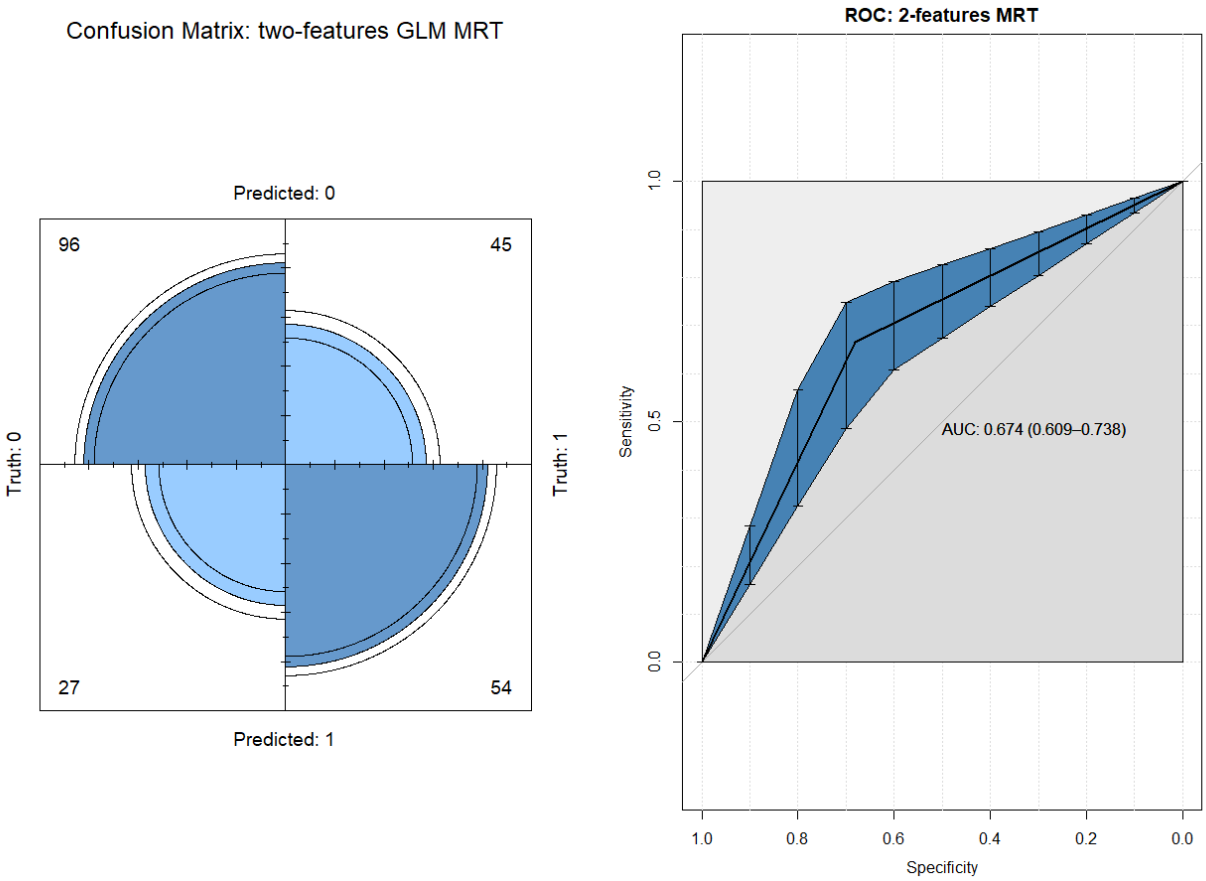

Fig 3.23 Confusion Matrix and ROC curve of two features MRT-GLM

2. PTT: In the case of PTT the lowest AIC was when only Cardinal Points was used as a feature (AIC: 300.11).
AIC's started to grow significantly only when the number of features was more than 5, in fact between the one feature model and the five feature model the ΔAIC was -0.70. Meanwhile the highest AIC's were found for full and almost full features models. For this model the AIC was around 320.

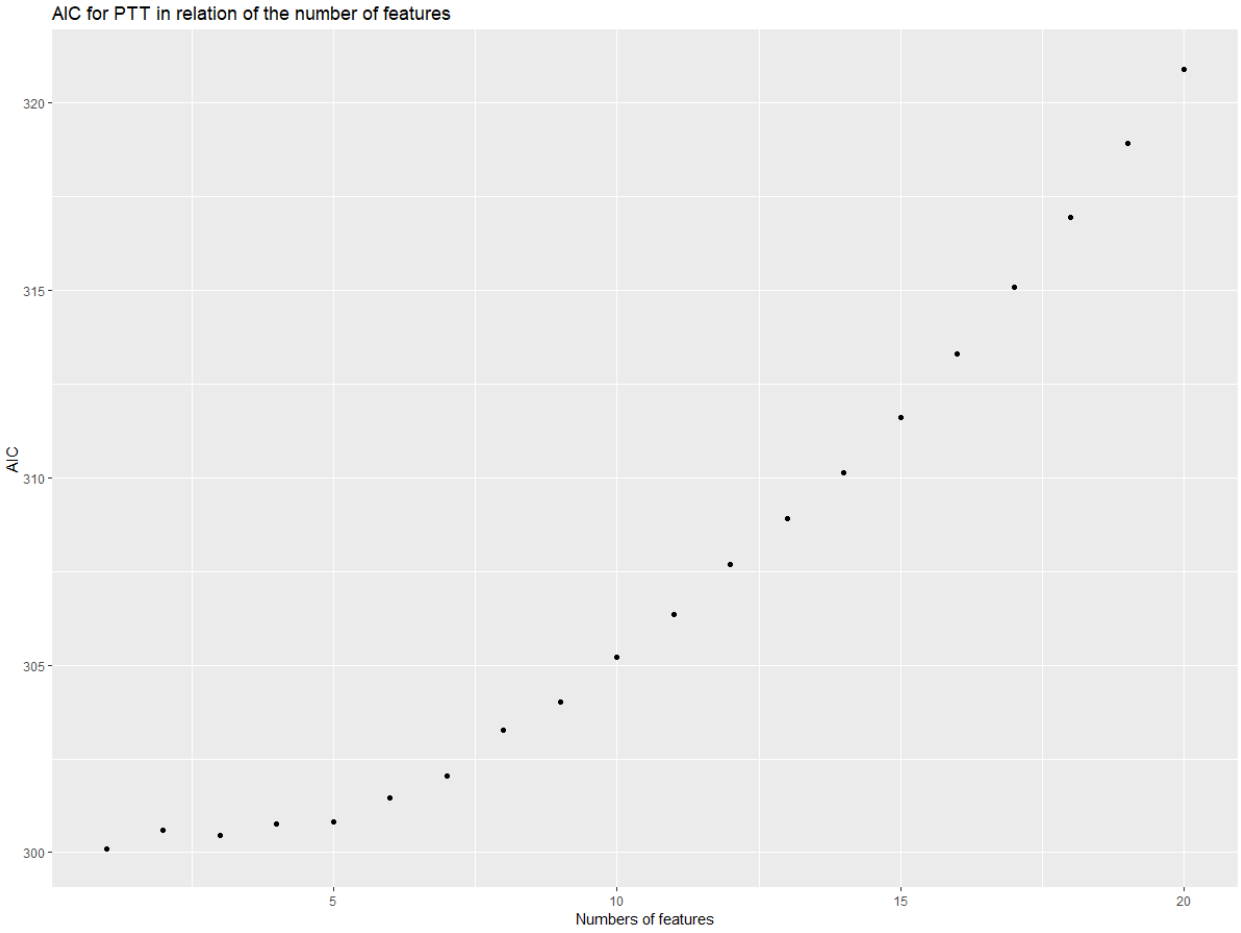

Fig 3.24 AIC's by numbers of features in MRT-GLM

From this previous observation we preferred to test a five features model. The features were: Cardinal Points, Emotion Control, Experience Opening, Perseverance and Culture Opening. The accuracy of this model grew up to 0.617 with a sensibility of 0.623, a specificity of 0.611 and an AUC of 0.617 (C.I. = 0.553-0.681).

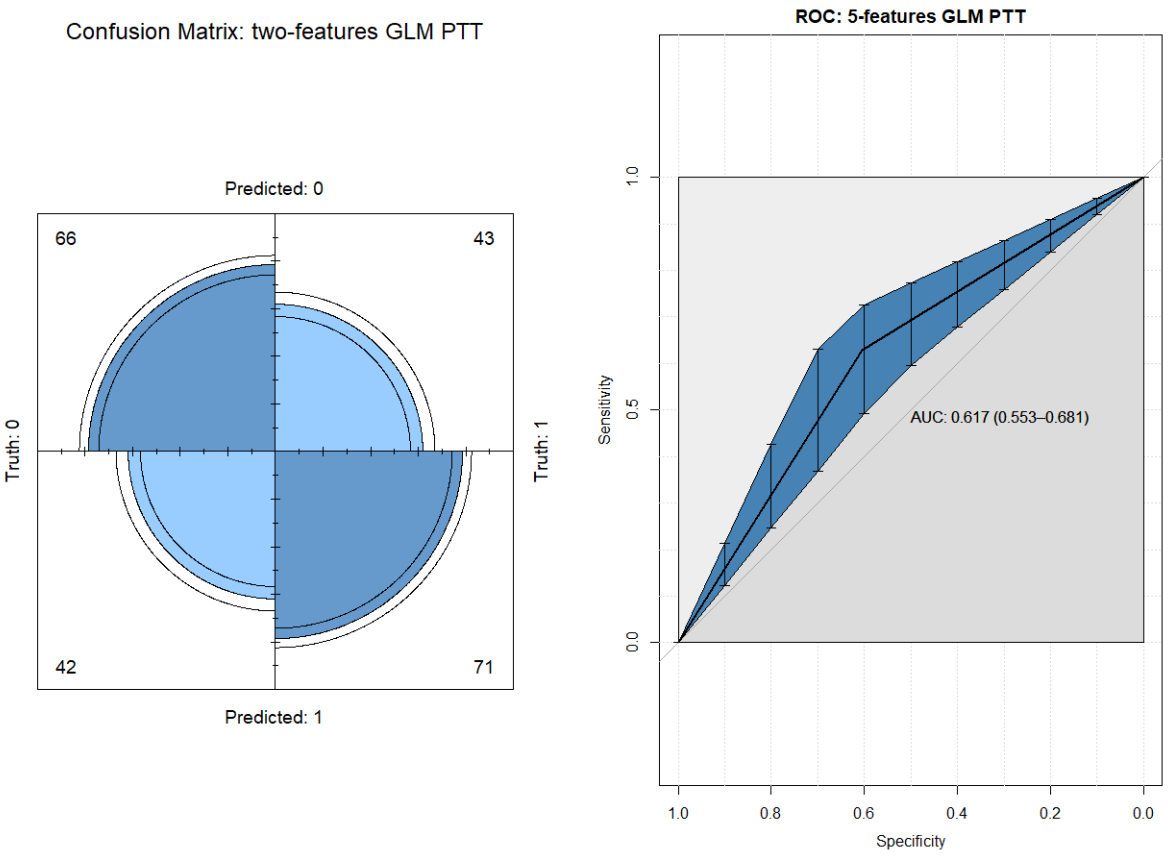

Fig 3.25 Confusion Matrix and ROC curve of two features PTT-GLM

3.5.2.3 Discussion of features selection

With this features selection procedures we managed to reach the highest accuracy so far for both MRT (accuracy 0.676) and PTT (accuracy 0.617). Moreover, we assessed the main role for Gender and Emotion Control to predict MRT, an effect found also in previous models (both LDA and QDA). For PTT Cardinal Points and other personality facets seemed relevant., differently by LDA and QDA models it disappear the effect of Gender, Map Use and Verbal Indication.

3.6 Polynomial Logistic Regression

After a standard logistic regression we tried to perform a logistic regression with polynomial of degree two.
Every features has been elevated to the square except for the variable Gender since elevating to the power of n a categorical variable is useless and misleading.
Result of full features model (features of degree one plus features of degree two), showed almost same results of a full features logistic regression model of degree one.

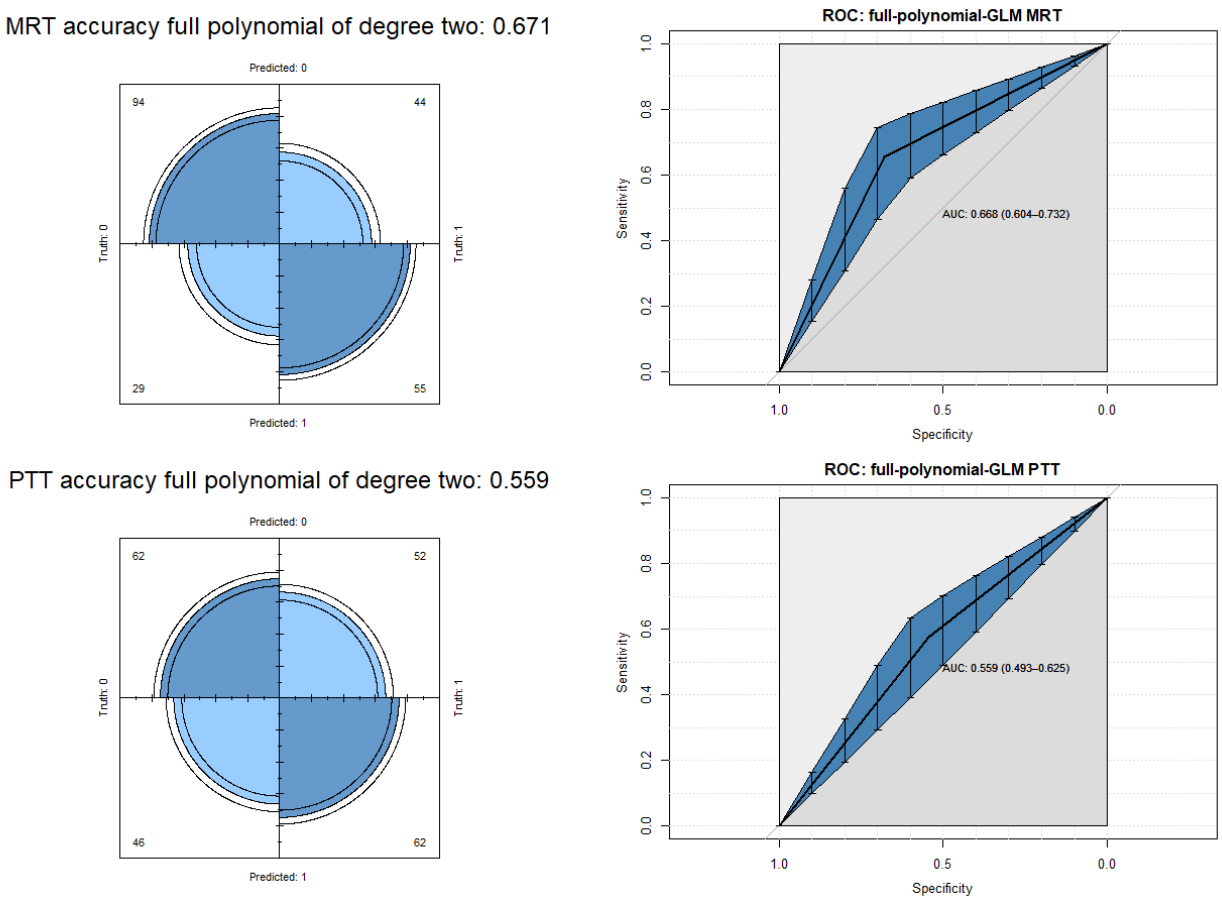

Fig 3.26 Confusion Matrix and ROC curve of polynomial GLM

3.6.1 Features Selection

The same procedure based on AIC, applied to the standard GLMs, has been used for features selection. We started from a 0 features model then we add the predictor which will create the model with the lowest AIC, we repeated this process for the 39 features.

1. MRT: the procedure showed that the models with the lowest AIC were the ones with 8-13 features. The best model was GLM with ten features (Gender, Emotion Control square, Cordiality, Cooperativity square, Cooperativity, Culture Opening square, Sense of Direction square, Sense of Direction, Experience Opening and Experience Opening square) and an AIC of 278.31. Meanwhile the worst model had 38 features and an AIC of 316.14.

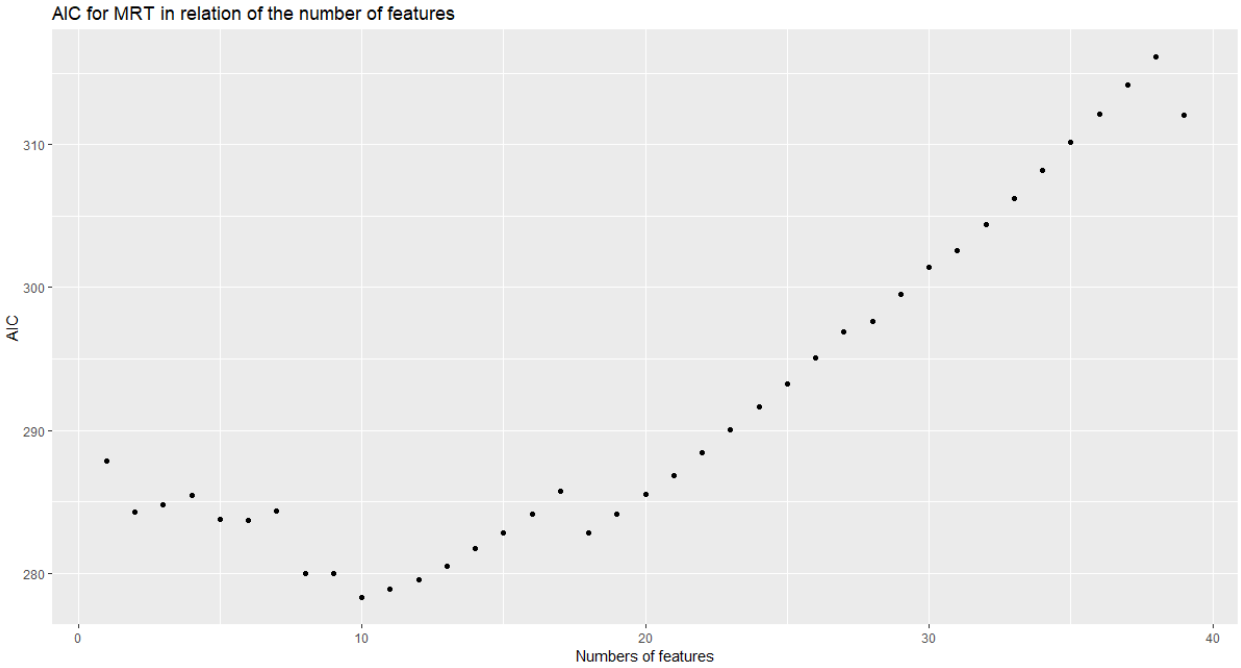

Fig 3.27 AIC based by features for a polynomial MRT-GLM

The result of these ten features GLM model showed discrete results and the best reached so far. The accuracy of the model was 0.739 with a sensibility of .636 and a specificity of .821, finally managing to get a significant sensibility. The AUC was 0.739 (C.I. = 0.680 - 0.799).

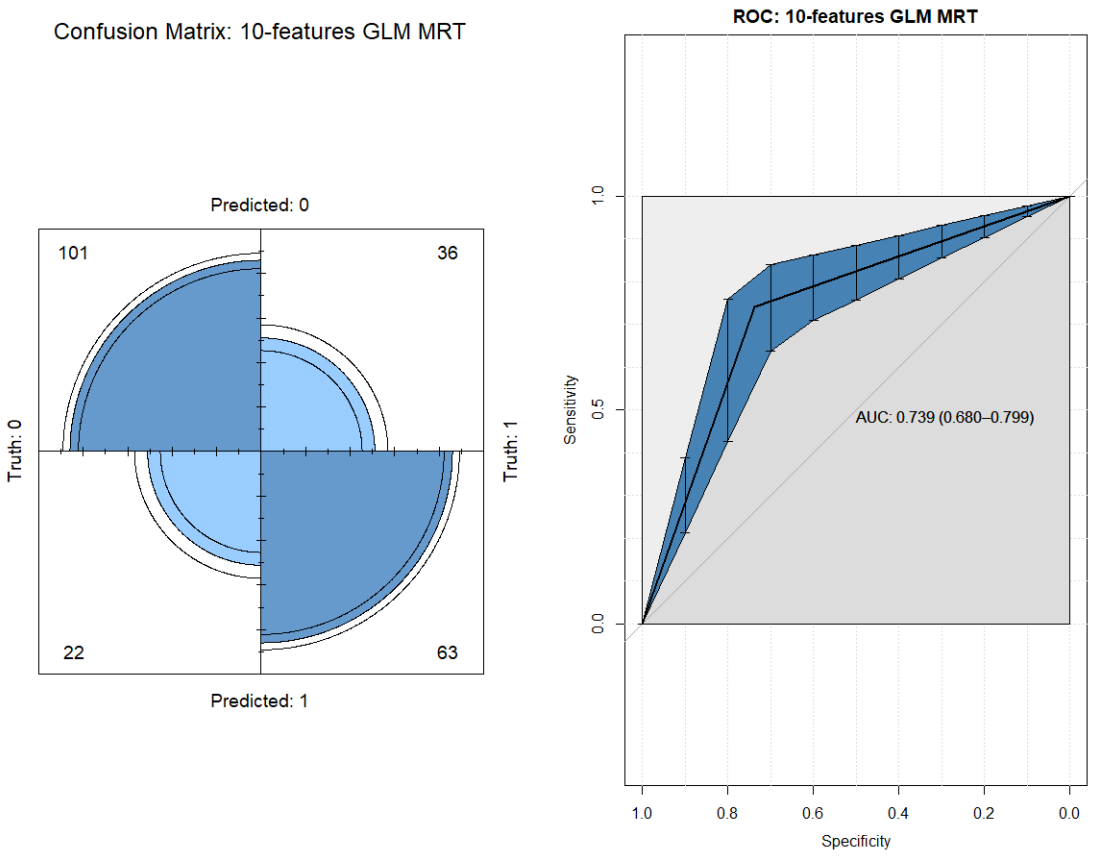

Fig 3.28 Confusion Matrix and ROC curve of a polynomial PTT-GLM

1. PTT: for the PTT, the models with the lowest AIC was the one with 14 features (Cardinal Point and Cardinal Points square, Cooperativity and Cooperativity square, Emotion Control and Emotion Control square, Experience Opening and Experience Opening square, Culture Opening and Culture Opening square, Map Use and Map use square, Perseverance, Cordiality square) and an AIC of 294.34. The worst model had 38 features and an AIC of 324.23.

The accuracy of the 14 features model was not as high as the previous PTT model: the accuracy was 0.604, the sensibility was 0.482 and the specificity was 0.732. The AUC was 0.614 (C.I. = 0.548 - 0.679).

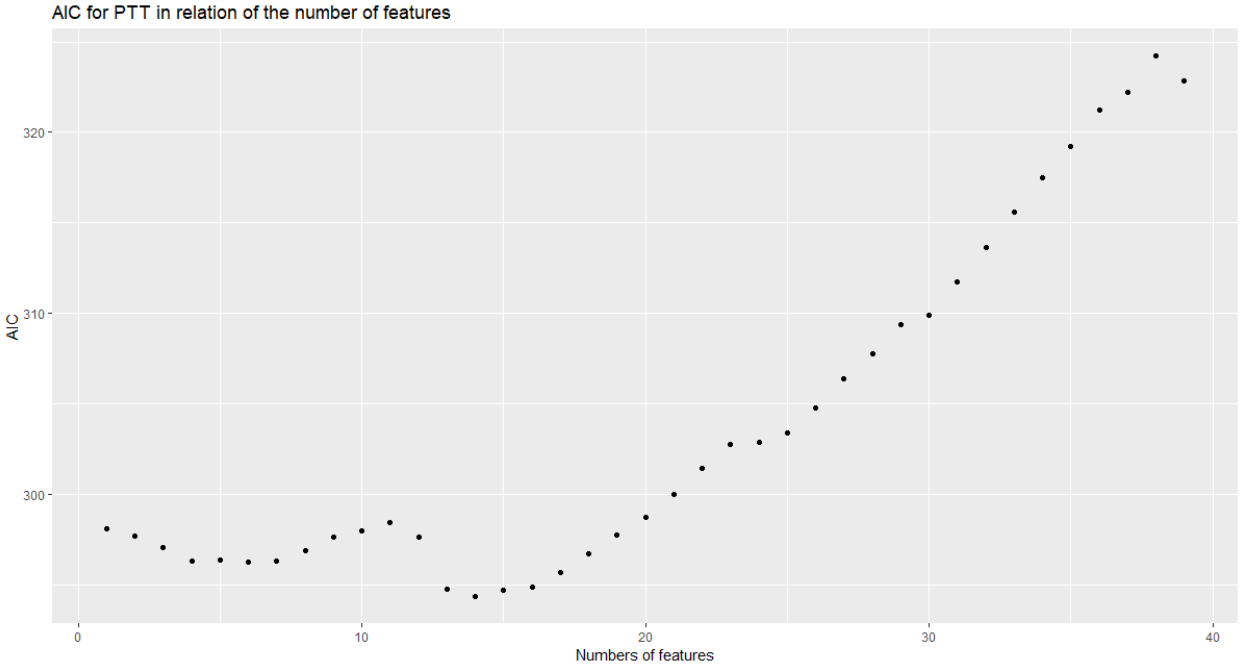

Fig 3.29 AIC based by features for a polynomial PTT-GLM

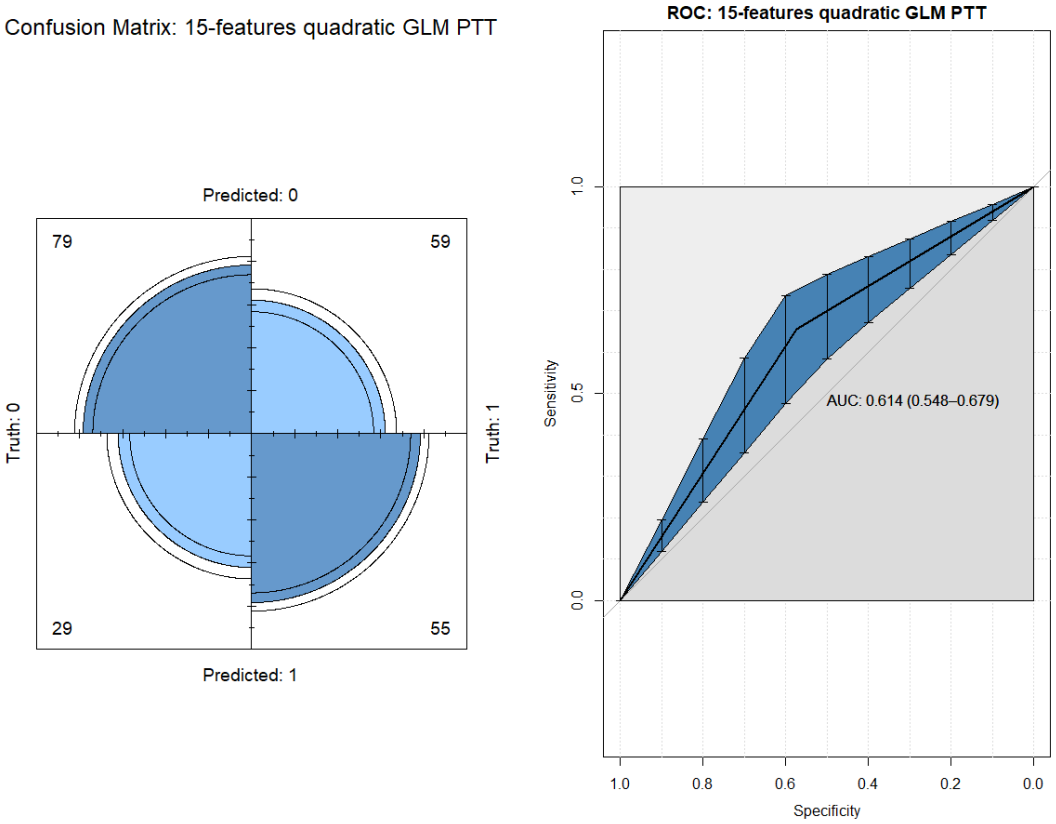

Fig 3.30 Confusion Matrix and ROC curve of a polynomial PTT-GLM

Chapter 4

Discussion

From this project we managed to extract meaningful information of the relation between Gender, Personality Traits, self-asessment wayfinding inclinations and Spatial Ability Tasks. In particular we tried to predict the binary scores (High: 1, Low: 0) of two Spatial Tasks: Mental Rotation Task and Point-perspective Taking Task.

Different algorithms for classification were used together with features selection procedures, we started with k-NN, followed by a Linear and a Quadratic Discriminant Analysis and finished with a Logistic Regression Model with polynomial of degree one and two.
In general we found that MRT was easier to predict rightfully than PTT, in fact the best model for MRT, a degree two polynomial GLM with 10 features, reached a 0.739 accuracy, meanwhile for PTT the best performing model, a degree one polynomial GLM with 5 features, reached only a 0.617 accuracy.

4.1 MRT results

The results, in terms of accuracy, in the case of MRT ranged from 0.653 to 0.739. The worst accuracy was found when QDA was applied meanwhile the best one was reached by using a polynomial logistic regression model.
All models showed a really low sensibility (0.434 - 0.636) and a discrete specificity (0.756-0.821), meaning a tendency to classify subjects towards the negative class/low spatial ability.

MRT	Features	Accuracy	Sensibility	Specificity	AUC
k-NN (k=18)	All	. 653	.434	.821	.656
LDA	6	.671	.526	.789	.670
QDA	2	.653	.526	.756	.649
GLM	2	.676	.535	.789	.674
GLM-poly	**10**	**.739**	**.636**	**.821**	**.739**

Table 4.1: *Results for the response variable MRT*

More surprisingly the best features were not the ones coming from the self-assessment wayfindings inclinations, as one would suspect, but they were Gender present as first features for all the models and Emotion Control present in all the model except QDA. There were also features such Sense of Direction and Cardinal points which were present in two of the four models.

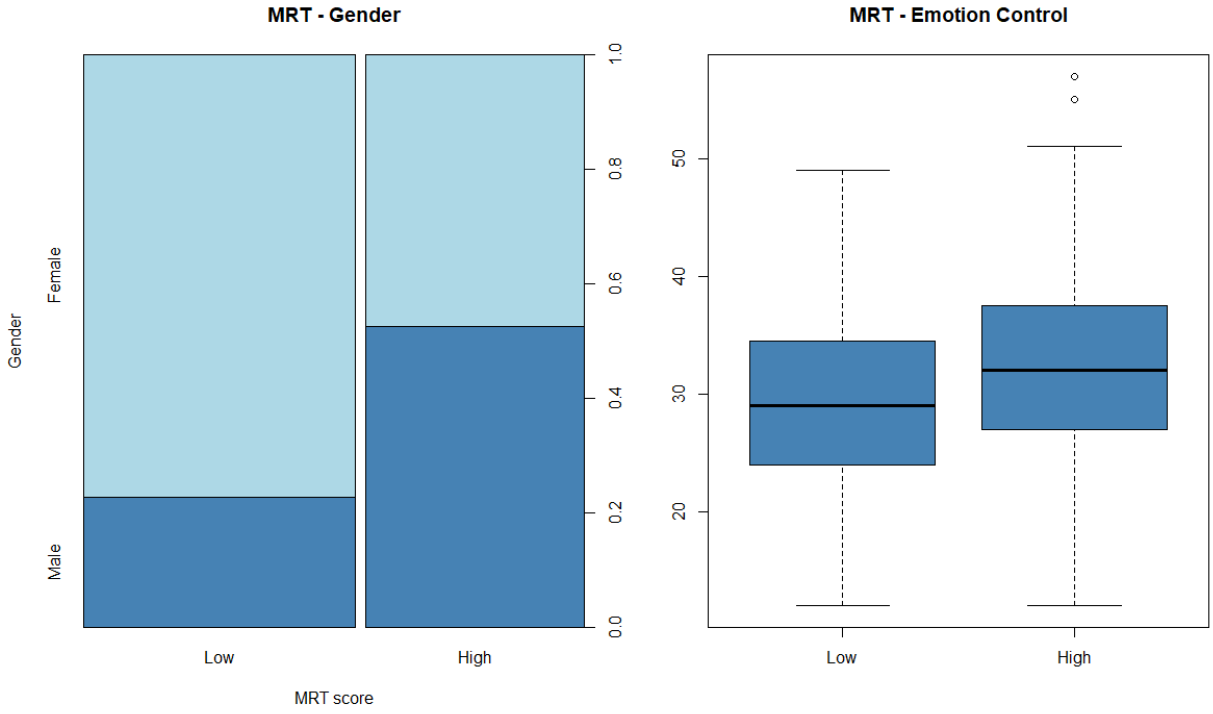

Fig 4.1 Relations between MRT and its main features

MRT	LDA	QDA	GLM	GLM-poly
1st	Gender	Gender	Gender	Gender
2nd	Emotion Control	Cardinal Points	Emotion Control	Emotion Control**2
3rd	Pulse Control	.	.	Cordiality
4th	QAS	.	.	Cooperativity**2
5th	Cardinal Points	.	.	Cooperativity
6th	Sense of Direction	.	.	Culture Opening**2
7th	.	.	.	Sense of Direction
8th	.	.	.	Sense of Direction**2
9th	.	.	.	Experience Opening
10th	.	.	.	Experience Opening**2

Table 4.2: *Features for the different MRT models*

4.2 PTT results

PTT models, as said before, were much worse than MRT models. The best PTT model was the degree one GLM with an accuracy of 0.617, while the worst was the QDA model with an accuracy of 0.559, which wasn't significantly higher than an accuracy of a model with No Information Rate.

PTT	Features	Accuracy	Sensibility	Specificity	AUC
k-NN (k=18)	All	.563	.500	.629	.556
LDA	4	.595	.579	.611	.595
QDA	4	.559*	.526	.593	.560
GLM	**5**	**.617**	**.623**	.611	**.617**
GLM-poly	14	.604	.483	**.735**	.614

Table 4.3: *Results for the response variable PTT*

By being present in all the four model, the main feature was for sure Cardinal Points. Other important features were Map Use, Emotion Control, Experience Opening and Culture Opening. Like in MRT models we observed a good effect of Emotion Control and Experience Opening.

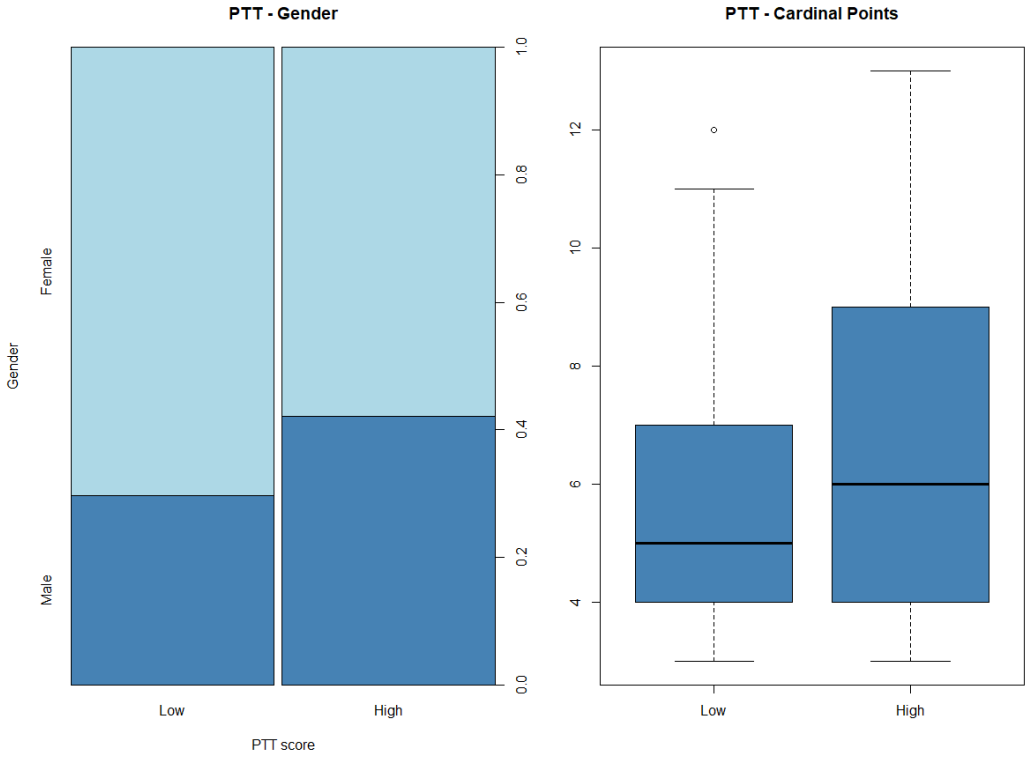

Fig 4.2 Relations between PTT and its main features

PTT	LDA	QDA	GLM	GLM-poly
1st	Gender	Gender	Cardinal Points	Cardinal Points**2
2nd	Map Use	Map Use	Emotion Control	Emotion Control**2
3rd	Verbal Indication	Verbal Indication	Experience Opening	Emotion Control
4th	Cardinal Points	Cardinal Points	Perseverance	Experience Opening
5th	.	.	Culture Opening	Perseverance
6th	.	.	.	Map Use**2
7th	.	.	.	Cardinal Points
8th	.	.	.	Map Use
9th	.	.	.	Culture Opening
10th	.	.	.	Culture Opening**2
11th	.	.	.	Cordiality**2
12th	.	.	.	Cooperativity**2
13th	.	.	.	Cooperativity
14th	.	.	.	Experience Opening**2

Table 4.4: *Features for the different PTT models*

4.3 Not what we expected. Why is that?

The results of the models and of the project are not surprisingly in terms of the accuracy reached, but more in terms of features extracted.
At the start of the project it seemed reasonable to expect high impact of the self-assessment wayfinding inclinations since those features are part of the spatial cognition field. But this is not quite what happened to result from our model: the features which appeared relevant for both MRT and PTT are Gender, Cardinal Points, Emotion Control (hugely present in all the models) and Experience Opening.
The effect of gender in spatial ability (Males tend to perform better) has been found multiple times in literature (Ben-Chaim, Lappan and Houng, 1988; Fennema and Tartre, 1985; Harris, 1981; Johnson and Meade, 1987) so it was expected to have Gender as a features in most of the models.
Cardinal Points as relevant feature can be reasonable, what is surprising is for it to be the only relevant features from all the wayfinding inclinations features.
The presence of Experience Opening could be justified by the fact that higher score in Experience Opening mean higher inclinations to go outside in new places which can stimulate our spatial reasoning. Sadly, this is hypothesis is refused since Experience Opening has a negative regression coefficients, showing that increase in Experience Opening decrease the odd ratio of success.
The most surprising feature is for sure Emotion Control: why should the capability of controlling my emotions determines my spatial ability? Probably this effect is found because MRT and PTT are two tasks which are taken under a really limited short time (no more than 5 minutes) and for this reason they can be considered under pressure tests. With them being under pressure tests the capability of a person to control his/her anxiety and his/her emotions can be hugely important in determining the score of MRT and PTT.
In conclusion we found that in order to predict MRT and PTT logistic regression models seems to work fine or at least better than LDA, QDA and that the best features are not really the wayfinding inclinations (except for Cardinal Points) but Gender, Emotion Control and Experience Opening.
Moreover from these results we exploit the fact that MRT and PTT could be not the best tasks to infer Spatial Ability since they showed strong relations with features such Emotion Control and weak relations with wayfinding inclinations, which should be expected to have the highest relations with spatial ability.

References

-Caprara, G. V., Barbaranelli, C., Borgogni, L. and Perugini, M. (2008). Big Five Questionnaire. Firenze: Organizzazioni Speciali.

-Munzer, S., Zimmer, H. D., Schwalm, M., Baus, J. and Aslan, I. (2006). Computer-assisted navigation and the acquisition of route and survey knowledge. Journal of Environmental Psychology, 26, 300-308. doi:10.1016/j.jenvp.2006.08.001.

-Lawton, C. A. (1994). Gender differences in way-finding strategies: Relationship to spatial ability and spatial anxiety. Sex Roles, 30, 765-779. doi:10.1007/BF01544230

-De Beni, R., Meneghetti, C., Fiore, F., Gava, L. and Borella, E. (2014). Batteria VS. AbilitÃă visuo-spaziali nell arco di vita adulta [VS Battery. Visuo-spatial abilities in the adult life span]. Firenze: Hogrefe.

-Pazzaglia, F.and Meneghetti, C., (2017). Acquiring spatial knowledge from different sources and perspectives: Abilities, strategies and representations. In J. M. Zacks, H. A. Taylor (Eds.). Representations in Mind and World. Essays Inspired by Barbara Tversky. Routledge, pp. 120-134.

-Vandenberg, S. G. and Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. Perceptual and Motor Skills, 47, 599-604, doi:10.2466/pms.1978.47.2.599.

-Kozhevnikov, M. and Hegarty, M. (2001). A dissociation between object manipulation, spatial ability and spatial orientation ability. Memory Cognition, 29, 745-756. doi:10.3758/BF03200477.

-Ben-Chaim, D., Lappan, G. and Houang, R. T. (1988). The Effect of Instruction on Spatial Visualization Skills of Middle School Boys and Girls. American Educational Research Journal, 25(1), 51-71. https://doi.org/10.3102/00028312025001051

-Fennema, E. and Tartre, L. (1985). The Use of Spatial Visualization in Mathematics by Girls and Boys. Journal for Research in Mathematics Education, 16(3), 184-206. doi:10.2307/748393

-Johnson, E. S. and Meade, A. C. (1987). Developmental patterns of spatial ability: An early sex difference. Child Development, 58(3), 725-740. http://dx.doi.org/10.2307/1130210