DALARNA UNIVERSITY

Data Mining - DT3019

Final Project Report – 08 November 2017

Mining Data of Executive Functions, Poker Skills and Gambling Problems

Author: Marina Ferreira Uchoa

h16mferr@du.se

1 Introduction

In this project data concerning 23 poker players was analyzed in order to investigate the relation between neuropsychological characteristics of a player and his/hers poker ability. Psychological tests were carried to assess problematic gambling and also to gather data with respect to each player's Executive Functions (EF) [1]. The resulting database is used in the present project to address the following questions.

1. Which poker variables are most important for judging poker skills?
2. Which EF variables are most related to which poker variables?
3. Which EF variables are best for separating strong from weak players, average from strong and average from weak?
4. Can poker variables predict problem gambling?
5. Can EF variables predict poker skills?
6. Can EF variables predict problem gambling?

Poker variables refer to summarized hand history of each player for 10 attributes such as cumulative amount won/lost and voluntary addition of money to the pot. Poker skills were assessed by three different poker experts, who analyzed the anonymized hand history and rated each player in a performance scale from 0 to 100.

EF variables refer to all the measures from the psychological tests performed to assess the Executive Functions. Problem gambling is assessed through three tests: Gambling Related Cognition Scale (GRCS-I), South Oaks Gambling Screen (SOGS) and Problem Gambling Severity Index (PGSI). These measures relate to different time spans: GRCS-I refers to the current time, PGSI to the previous year and SOGS to the individual's whole life.

2 Methodology

Questions 1 through 3 are related to knowledge extraction. To answer these questions, correlation and association analysis were used for variable selection and generalized linear models for model building. When possible, linear regression was preferred for its simplicity. However, when the response variable was categorical, logistic regression was used to perform classification.

For questions 4 through 6, several data mining techniques were considered: generalized linear model, decision tree (CART), random forest, bagged CART, XGBoost, Support Vector Machine (SVM) and k-Nearest Neighbour (kNN).

A leave-one-out cross validation was performed. Since there are few observations, this approach ensures maximum training data size while not compromising the predictive performance. All possible combinations of training and testing are covered and the models' performance can be assessed by checking the average squared root mean error (RMSE). It is important to note that due to the nature of this procedure, categorical variables with

only one observation in a given class had to be removed from the analysis - they would be constant for when the observation left out was the one observation in that category.

Since `rattle` does not provide in-built tools for regression or non-binary classification [2], models were built directly in R. The regressions were performed with the original values for the gambling scores and average rating of the players.

In order to tune the models, anova analysis was used for variable selection in linear regression, a grid search was performed for SVM and the function `train` of the `caret` package was used on the remaining models [3]. This function automatically performs a grid search over the parameters of the proposed model, returning the best set of parameters according to the inputted selection criteria - I used minimization of the RMSE.

3 Data Preprocessing

In this phase, the data set was cleaned and structured to allow further analysis. Unrecognized or undesired symbols were removed, categorical variables were set as factors and the identifiers for all observations were standardized as the poker mapper identifier (PKMP.ID). The observations in poker variables were reduced to unique entries for each player. The data kept was that when the player played the most hands, since it is assumed that it better represents the player's performance. The records kept are for players that had skills, hand history and psychological tests data.

The final data for poker variables contains data related to 23 players with various ability levels. There are 10 players with average rating less than 40 (weak), 8 with rating between 40 and 70 (average) and 5 with rating 70 or over (strong). With regard to gambling problems, based on SOGS, there are 10 non-problematic players, 12 problematic and 1 pathological, while according to PGSI there are 2 non-problematic, 1 with low level of problems, 20 with medium level and no one with really problematic gambling issues.

With regard to EF and gambling variables, those with constant values for all players, as well as those with very high correlation (above or equal to 90%) were removed. This resulted in 302 EF variables and 8 gambling related variables, all with 23 observations.

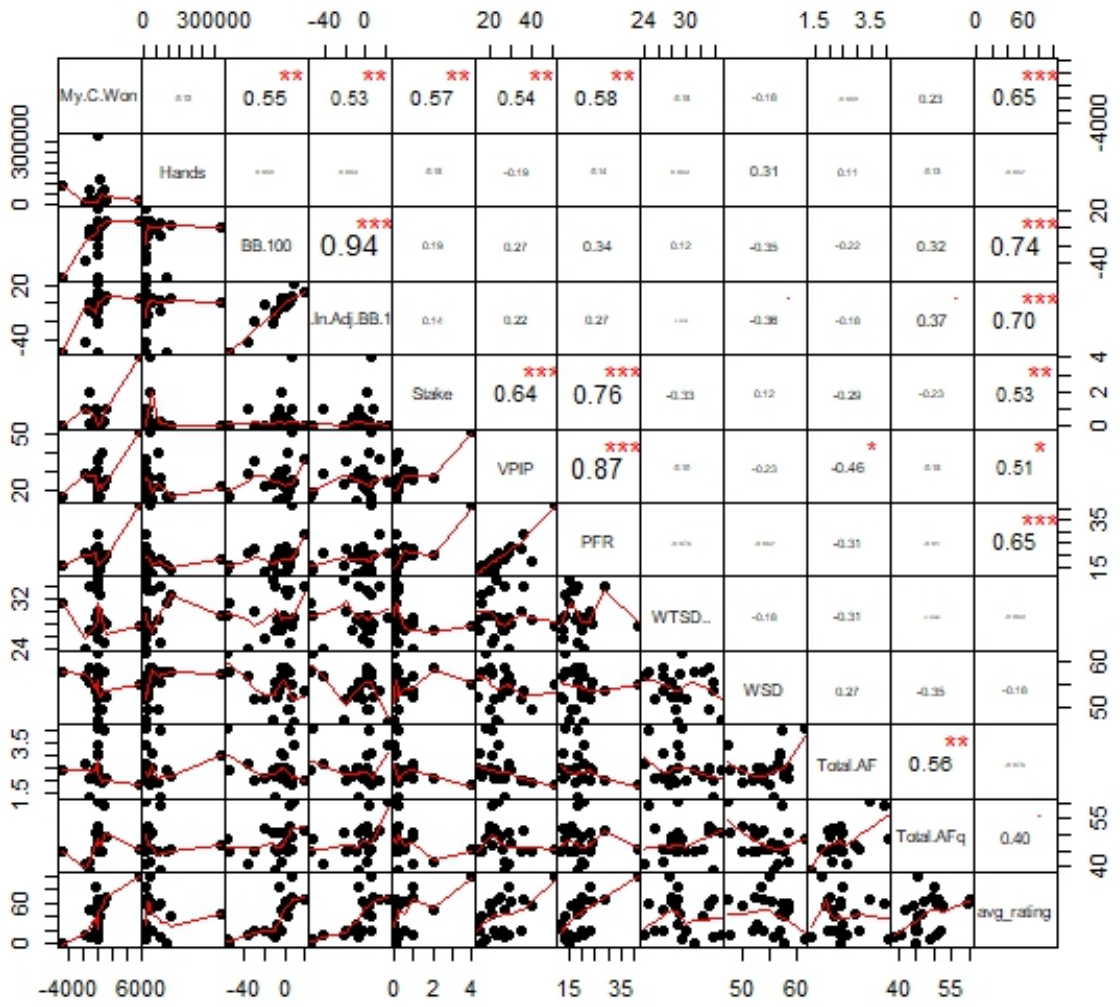
4 Data Analysis

4.1 Which poker variables are most important for judging poker skills?

To answer this question, the correlation between all poker variables and the average rating of each player was calculated. Figure 1 plots the correlations in the upper triangle with the font size proportional to the strength of the correlation. Variables My.C.Won, BB.100, All.In.Adj.BB.100, Stake, VPIP and PFR have correlations with average rating of magnitude greater than 0.5, while Total.AFq has a correlation of 0.40. However, many of these are highly correlated among themselves. To decide with regard to which are most relevant, a second analysis was carried out.

To complement the correlation analysis, a linear regression was put through. Initially, all variables were included and at each consecutive step the variable with the highest p-value from an anova chi-squared test was removed until all remaining variables were

Figure 1: Correlation of Poker Variables and Average Rating



significant to the 20% level, as suggested by [4]. This model kept both Total.AFq and Total.AF, however with a non-significant estimate for the latter and for My.C.Won. Further analysis with a chi-squared test proved that removing these variables positively impacted the performance of the model. The estimates and summary statistics of the final model are displayed in Table 1.

Table 1: Linear Model between Poker Variables and Average Rating

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	-56.5570	29.4841	-1.92	0.0711
BB.100	0.6929	0.1546	4.48	0.0003
Stake	10.4112	4.8834	2.13	0.0470
PFR	0.9307	0.6857	1.36	0.1915
Total.AFq	1.6941	0.5825	2.91	0.0094

4.2 Which EF variables are most related to which poker variables?

To answer this question, from all EF variables, those which are numeric and have a correlation higher than 0.7 with any of the poker variables and those which are categorical and have an intraclass correlation (ICC) higher than 0.75 were considered to be most related with the poker variables. The results are available in Table 2.

Table 2: Correlation and ICC between EF and Poker variables

	EF_variable	Pk_variable	Correlation
5	TOL_2_TOT_FA	My.C.Won	0.72
4	TOL_3_TOT_TSEC_TOT	Hands	0.73
3	TOL_2_TOT_MOVE_2	Stake	0.74
2	TOL_2_TOT_FA	PFR	0.76
1	TOL_2_TOT_FA	Stake	0.85
	EF variable	Poker variable	ICC
12	EQIS_SMEQ_VLW_70_79	All.In.Adj.BB.100	0.77
11	EQIS__PIEQ_VLW_70_79	All.In.Adj.BB.100	0.78
10	RSBQ_P_50._300_LOSS_LEVG	Total.AFq	0.78
9	EQIS_INTRA_HI_110_119	Total.AFq	0.78
8	RSBQ_P_100_CERTLOSS	Total.AFq	0.78
7	EQIS__PIEQ_VLW_70_79	BB.100	0.79
6	EQIS_SMEQ_VLW_70_79	Total.AF	0.79
5	EQIS_SMEQ_VLW_70_79	BB.100	0.79
4	EQIS_GMEQ_VHI_120_129	VPIP	0.86
3	EQIS__PIEQ_VLW_70_79	My.C.Won	0.87
2	EQIS_TOTEQ_LOW_80_89	All.In.Adj.BB.100	0.87
1	EQIS_TOTEQ_LOW_80_89	BB.100	0.89

From this result we have that Emotional Quotient Inventory Short (EQIS), which evaluates emotional-social intelligence with respect to well-being and success in life, and Tower of London (TOL), which measures the goal setting ability of the player, are the tests that individually have the highest impact on the poker variables. These are followed by Fredericks Risk Seeking Behaviour Questionnaire (RSBQ), which assesses the visual-spacial, planning and praxis abilities.

4.3 Which EF variables are best for separating strong from weak players, average from strong and average from weak?

For this task, a two step analysis was performed. First the relation between EF variables and the poker ability for each reduced subset was computed. For categorical variable pairs, Cramer's V measure of association was computed, while for categorical-numerical pairs, the ICC was calculated.

The idea was to select those variables with a relation above a certain threshold in order to perform a logistic regression. A model with over 300 variables and only 15 observations for analyzing strong and weak, 13 for average and strong and 18 for average and weak would be too complicated and possibily redundant. The threshold was adjusted to ensure that at least 5 variables were selected for each analysis.

For separating strong and weak players, a threshold of 50% for both Cramer's V and ICC was used. Whereas, for differentiating average and strong players, the threshold was of 40% and for average and weak, a minimum of 45% was used. Then, in model building, anova chi-squared tests were used to select which variables to keep in the model. The results for all groups are displayed in Table 3.

Table 3: Variables that Best Separate the Level Groups

Strong & Weak	Strong & Average
FLUS_FR_TOT_ER	TOL_7_TOT_TSEC_EXE
IBQ_N_HOW_MUCH_IMPULSIVE_.5_.5	DIGIT_FW_MAXSEQ_0_9
TOL_7_TOT_TSEC_EXE	TOL_8_TOT_MOVE_4
Average & Weak	
IBQ_N_HOW_MUCH_IMPULSIVE_.5_.5	
TOL_2_TOT_TSEC_EXE	
MAX_DAILYBET_100_1.000EUR	

By analyzing the relevant variables for each binary group level, it is perceivable that the TOL test is important for all groups. Frederick's Intertemporal Behaviour Task (IBQ), which measures impulsivity, had the same feature relevant in two groups. DIGIT, at its term, measures short-term memory and has a connection with separating strong from average players. While the Semantic Verbal Fluency Test (FLUS) was relevant for distinguishing strong and weak players. Additionally, one variable from the general characteristics of the player was deemed relevant, namely whether a player has a maximum daily bet between 100 and 1000 euros.

4.4 Can poker variables predict problem gambling?

Three analyses were carried out each using one of the gambling problem related tests - GRCS, PGSI and SOGS - with each of the proposed methods. For each model, a leave-one-out cross validation was performed and the mean root mean squared error (RMSE) was computed as a measure of model predictive performance. For linear regression, the variables selected whereas for the other methods, the tuned parameters are displayed along with the corresponding RMSE in Table 4.

Table 4: Models for Poker and Gambling Variables

GRCS

Model	Parameters	RMSE
Linear Regression	Null model	18.3
SVM	$\epsilon = 0.0152$, cost = 12	18.5
CART	Complexity parameter = 0.128	18.2
Random Forest	Number of variables to possibly split at in each node = 2, split rule: extratrees	19
Bagged CART		20.2
XGBoost	Boosting iterations = 50, maximum tree depth : 2, $\eta = 0.3$, $\gamma = 0$, proportion of column sample by tree = 0.8, minimum child weight = 1, percentage subsample = 0.5	22.7
kNN	k = 7	18.9

PGSI

Model	Parameters	RMSE
Linear Regression	All.In.Adj.BB.100, VPIP, PFR, WSD, Total.AFq	2.02
SVM	$\epsilon = 0.00525$, cost = 89	1.82
CART	Complexity parameter = 0.0834	1.99
Random Forest	Number of variables to possibly split at in each node = 2, split rule: variance	1.88
Bagged CART		1.45
XGBoost	Boosting iterations = 50, maximum tree depth : 1, $\eta = 0.3$, $\gamma = 0$, proportion of column sample by tree = 0.8, minimum child weight = 1, percentage subsample = 1	1.67
kNN	k = 9	1.94

SOGS

Model	Parameters	RMSE
Linear Regression	VPIP	1.29
SVM	$\epsilon = 0.0109$, cost = 67	1.9
CART	Complexity parameter = 0.137	1.45
Random Forest	Number of variables to possibly split at in each node = 2, split rule: extratrees	1.41
Bagged CART		1.23
XGBoost	Boosting iterations = 150, maximum tree depth = 1, $\eta = 0.3$, $\gamma = 0$, proportion of column sample by tree = 0.8, minimum child weight = 1, percentage subsample = 1	1.61
kNN	k = 9	1.41

The first thing to note is that the reported results are steadily comparable for all models except the linear model. Due to nature of linear regression, only variables with *linear* will be deemed relevant. Therefore, the presented RMSE is for the "tuned" linear regression, i.e. after a second phase of variable selection. Since the other models can handle non-linear relationships, all variables were kept for modelling.

To analyze the results in Table 4, it is also important to know the possible range for the target variable of each test. GRCS has a total score ranging from 23 to 161, therefore the lowest RMSE (18.2 for CART) represents around 13.2% of its scope. It is therefore, a relatively high error rate. Additionally, the linear model found no variable to have a relation with GRCS' total score.

For PGSI, the range of the target variable is from 0 to 27. The lowest RMSE (1.45 for the bagged CART) represents only 5.3% of its range, meaning there is a much more reliable estimation than for GRCS. Finally, SOGS' total score ranges from 0 to 20 and the lowest RMSE (1.23 also for bagged CART) is equivalent to 6.15% of its domain.

The presented results leads to believe that predictions based on poker variables are possible and somewhat reliable for PGSI and SOGS, but not so reliable for GRCS.

4.5 Can EF variables predict poker skills?

To answer this question, the first step was to create a reduced data set with only the most relevant variables. Numeric variables with correlation higher than 0.45 with the average rating, as well as categorical variables with ICC higher than 0.45 were kept for modelling. Then, the proposed methods were computed and parameters tuned, yielding the outputs in Table 5.

Table 5: Models for EF and Poker Skills Variables

Model	Parameters	RMSE
Linear Regression	SPM_D_T_SEC, FLUS_FR_TOT_ER, EQIS_INTER_EQ_STD_50_150, EQIS_SM_EQ_STD_50_150	18.3
SVM	$\epsilon = 0.00323$, cost = 1	17.9
CART	Complexity parameter = 0.19	31.9
Random Forest	Number of variables to possibly split at in each node = 6, split rule: extratrees	19.2
Bagged CART		22.5
XGBoost	Boosting iterations = 50, maximum tree depth : 1, $\eta = 0.3$, $\gamma = 0$, proportion of column sample by tree = 0.8, minimum child weight = 1, percentage subsample = 0.75	16.9
kNN	k = 5	23.9

AS mentioned before, to analyze the predictive ability of the models, it necessary to take into account the range of the target variable. In this case, the target was average rating ranging from 0 to 100. Hence, an RMSE of 16.9 means that the predicted rating from the XGBoost model was on average 16.9 points far from the actual value.

Taking into account the classification into weak (0-39), average (40-69) and strong (70-100) players, players with skill predicted between 53.1 and 55.9, could in reality be in

any of the proposed categories. This overlap, however small, supports that poker skills cannot be reliably predicted using EF variables.

4.6 Can EF variables predict problem gambling?

For each test, the EF variables with correlation or ICC higher than 0.45 were kept for modelling. As in Subsection 4.4, the proposed models were applied to each test, with the target being their total scores. The results are presented in Table 7.

Beginning with GRCS, the lowest RMSE was achieved by the linear regression (13.7) and represents 9.9% of its total scope. This is still a high error rate, even if much smaller than when using poker variables to predict problem gambling (p.6). For PGSI, the lowest RMSE was obtained by XGBoost (1.3). This error represents 4.8% of the range of the target variable. Finally, the lowest RMSE when predicting SOGS' total score was of 0.728 by the linear model. This is equal to only 3.6% of the total range of the variable, indicating a very good accuracy.

To discuss the fact that the linear model had the best or second best RMSE for all tests, it is worth noting that even though its results were based on a further reduced data set whereas the other models were not, the RMSE of the other models with this further reduced data set was even higher.

As an experiment, the proposed methods were ran for GRCS and, with the exception of the decision tree, they performed the same or worse when the data set was smaller (see Table 6). This is likely due to the fact that those models that can also capture non-linear relationships had valuable data removed by removing the variables not used in the linear model, whereas for the decision tree, the further reduced data set was less complex than the bigger one.

Table 6: Models for EF and GRCS with Variables Selected by the Linear Model

Model	Parameters	RMSE
SVM	$\epsilon = 0.00646$, cost = 1	15.7
CART	Complexity parameter = 0.0892	18.8
Random Forest	Number of variables to possibly split at in each node = 2, split rule: extratrees	17.5
Bagged CART		17.9
XGBoost	Boosting iterations = 50, maximum tree depth : 1, $\eta = 0.4$, $\gamma = 0$, proportion of column sample by tree = 0.6, minimum child weight = 1, percentage subsample = 0.5	16.3
kNN	k = 9	16.7

The findings suggest that all three gambling problem assessment tests can be predicted with the results from EF tests, even though with different expected error rates. GRCS is still the most difficult to predict.

Table 7: Models for EF and Gambling Variables

GRCS

Model	Parameters	RMSE
Linear Regression	FLUF_ALLER_RIP, IBQ_L_PAY_OVERNIGHT_SHIPPING	13.7
SVM	$\epsilon = 0.00162$, cost = 1	14.9
CART	Complexity parameter = 0.178	20.1
Random Forest	Number of variables to possibly split at in each node = 4, split rule: extratrees	16.3
Bagged CART		17.8
XGBoost	Boosting iterations = 50, maximum tree depth : 3, $\eta = 0.4$, $\gamma = 0$, proportion of column sample by tree = 0.6, minimum child weight = 1, percentage subsample = 0.5	14.8
kNN	k = 7	16.6

PGSI

Model	Parameters	RMSE
Linear Regression	MAX_DAILYBET_100_1.000EUR, TOL_5_TOT_TSEC_EXE, SPM_D_T_SEC	1.32
SVM	$\epsilon = 0.00182$, cost = 34	1.43
CART	Complexity parameter = 0.213	1.52
Random Forest	Number of variables to possibly split at in each node = 8, split rule: extratrees	1.4
Bagged CART		1.53
XGBoost	Boosting iterations = 150, maximum tree depth : 3, $\eta = 0.4$, $\gamma = 0$, proportion of column sample by tree = 0.6, minimum child weight = 1, percentage subsample = 1	1.3
kNN	k = 7	1.55

SOGS

Model	Parameters	RMSE
Linear Regression	SPM_A_COR_0_12, SPM_B_COR_0_12, FLUF_L_ER_VAR, RSBQ_D_100_CERTGAIN, EQIS_SMEQ_HI_110_119, EQIS__MED_90_109	0.728
SVM	$\epsilon = 0.00141$, cost = 23	1.1
CART	Complexity parameter = 0	1.3
Random Forest	Number of variables to possibly split at in each node = 2, split rule: extratrees	0.98
Bagged CART		1.23
XGBoost	Boosting iterations = 50, maximum tree depth : 3, $\eta = 0.3$, $\gamma = 0$, proportion of column sample by tree = 0.6, minimum child weight = 1, percentage subsample = 0.5	0.835
kNN	k = 5	1.14

5 Final Considerations

This study encompassed analyzing data for 23 poker players with regard to their skills, executive functions and problem gambling assessments. The findings suggest that EF variables can be used for reliably predicting problem gambling but not poker skills. Additionally, poker variables can also predict problem gambling, but with worse accuracy.

In the modelling phase, for both gambling test results and average rating, the numerical target was kept. Different and perhaps better results could be found by classifying the players into categories such as weak, average and strong for poker skill predictions and as non-problematic, problematic and pathological for SOGS or as non-problematic, low level of problems, medium level and problematic for PGSI, for example. A possible approach for such problems would be to conduct multinomial logistic regression. However, it is a personal limitation that I do not know neither how to program not how to analyze the results of such regression.

An attempt to use Artificial Neural Networks (ANN) was also contemplated. However, the results were suspiciously the same RMSE for all parameter values tested. Having tried to solve this issue without success, I opted to withdraw any mentions of ANN in this report and in the code handed in with it.

Given the time and knowledge constraints, this project was executed as the most broad possible. A personal reflection is that the project challenged me both in terms of understanding the algorithms and in terms of programming.

References

[1] Mauro Schiavella, Matteo Maria Pelagatti, Jerker Westin, Gabriele Lepore, and Paolo Cherubini. Profiling online poker players. are executive functions correlated with poker ability and problem gambling? *Journal of Gambling Studies*.

[2] Graham J. Williams. *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. Use R! Springer, 2011.

[3] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. *caret: Classification and Regression Training*, 2017. R package version 6.0-77.

[4] U. Olsson. *Generalized Linear Models: An Applied Approach*. Lightning Source, 2002.