02445 - EVALUATING PREDICTIVE MODELS (TASK 2)

INDIVIDUAL ASSIGNMENT

Lucas Rieneck Gottfried Pedersen
s234842

June 21, 2024

ABSTRACT

This assignment investigates the feasibility of using heart rate data to predict self-reported frustration levels in an experimental setup. Utilizing a dataset comprising summary statistics of heart rate measurements from two cohorts, five predictive models were developed and evaluated: Artificial Neural Network (ANN), Random Forest (RF), k-Nearest Neighbors (KNN), Ridge Regression, and Logistic Regression. Each model was assessed using 8-fold cross-validation, with Mean Squared Error (MSE) as the performance metric. Despite extensive hyperparameter optimization and model comparison, results indicate that heart rate data alone does not reliably predict frustration levels, as no model outperformed the baseline significantly. However, it was not possible to prove any statistical differences when including the variability of the training set. The study discusses potential areas for further research, including the exploration of raw heart rate data and additional model variants to improve predictive accuracy.

1 Dataset and purpose

In this assignment, the feasibility of using only heart rate data to predict the self-reported frustration level of people in an experiment is investigated. The data is collected from the article Sneha Das [2024]. The exact features used for prediction are summary statistics of a 5 minute interval of heart rate data, resulting in 6 features:

- HR_Mean: mean heart rate. Continuous, ratio
- HR_Median: median heart rate. Continuous, ratio
- HR_std: the standard deviation of the heart rate. Continuous, ratio
- HR_Min: the smallest heart rate observed. Continuous, ratio
- HR_Max: the largest heart rate observed. Continuous, ratio
- HR_AUC: The AUC (Area under curve) of the heart rate. Continuous, ratio

The target we are trying to predict is "Frustrated" (0-10). It is discrete and ratio. Besides the aforementioned features, the dataset also contains

- Cohort: a variable indicating which cohort the datapoint is from. The experiment was conducted 3 separate times, leading to 3 cohorts. Only cohort 1 & 2 is used in this report. Discrete, nominal
- Individual (1-14): an index for the individual. Discrete, nominal
- Puzzler: a binary variable describing the participant's role (puzzler/solver). Binary, nominal
- Round (1-4): the round number, as each individual goes through 4 rounds of puzzle-solving. Discrete, nominal
- Phase (1-3): the phase number, indicating whether the data is collected from the pre-puzzle phase, the puzzle phase, or the post-puzzle phase. Discrete, nominal

Briefly described, the individuals in the experiment were put in groups of two, they were assigned puzzler/solver respectively, and competed against other groups in solving tangrams in 4 different rounds. Each round consists of three 5-minute phases, and in the pre- and post-puzzle phase, the participants were asked to rest. In the puzzle phase they solved puzzles. After each phase, the participants filled out questionnaires, and it is here they rated their frustration, among other emotions.

2 Methods

2.1 Models

In this assignment, 5 models were developed and compared. All models were trained on standardized data (standardized on training data), and the full 6 features were used for all models. This is due to the fact that when making scatterplots of the input features against the frustrated variable, there wasn't any visible correlation for any of them, and thus it wasn't discernible which features to remove. Backward/forward selection would be ideal to use instead, but took too many resources to perform in the scope of this assignment. The 5 models are:

- ANN: An artificial neural network with the structure 6-512-256-128-64-1. Implemented with early stopping on hold-out test set (10% of the training data), Adam optimizer, ReLU activation function, Kaiming normal weight initialization and with 3 independent trainings to choose the model with best test error. By far the most complex model with its 176.129 parameters, however, with early stopping the complexity is reduced.
- RF: A random forests model, using crossvalidation to find the optimal number of trees between 1 and 250 with a stride of 5 (i.e. tested 1-6-11-16 etc. no. of trees).
- KNN: A k-nearest-neighbor model, using crossvalidation to find optimal k between 1 and the amount of datapoints in the training set.
- Ridge regression, a linear regression model with ridge regularization. The regularization constant was found with crossvalidation, testing values between 10^{-3} and 10^{6}
- Logistic Regression (LR): A logistic regression model with L2 regularization, where the regularization strength (C) is tested over a range from 10^{-3} to 10^{3}. The logistic regression model is unusual for a task in predicting a ratio label. It was introduced to examine whether this task would benefit from a classification approach.
- Baseline: Simply the mean of the training set

Round/cohort	1	2
1	21	18
2	21	18
3	21	18
4	21	18

Table 1: Example of one of the outer folds used. The green cells indicate data used for training, the red cell indicates data used for validating, and the gray cells indicate data that were not used. Numbers indicate how many datapoints were in each cell. Note that since the individuals differed between cohorts, it is also made sure we don't test on individuals that were trained on.

For detailed information on the exact training methods, we encourage readers to refer to the code & data available on GitHub[1]. The models were chosen to ensure a wide spectrum of different model complexities and types. In the exploratory phase of this study, it became very apparent that the target was challenging to predict based on the features. Therefore, to give the best possibility of prediction, many types of models were tested. Note, for hyperparameter optimization, we performed cross-validation on the training set to create test sets. The final validation set remained wholly untouched until we evaluated each of the six final models (per outer fold).

2.2 Crossvalidation

To assess the performance of our different models we use 8-fold crossvalidation. The data consists of different groups (phase, puzzler, round, cohort, individual), which could introduce some covariance in the dataset. To minimize the effects of this, one usually makes sure not to split the groups between testing/training sets. However, when seperating all groups, there are only 18-24 datapoints left for training, which is low. Therefore, we have chosen to separate only individual, cohort and round between testing and training. An example of this is shown in table 1.

We separated cohorts because it is shown in Sneha Das [2024] that the data for some of the measurements varied significantly between cohorts. As we want to be able to predict data that happens on other days, we therefore only test on data from another cohort. We separate rounds, because, assuming the rounds are setup up the same way across individuals, some rounds may contain more frustrating puzzles than others, and this could technically introduce some specific patterns in the correlation between heart rate and frustration level. Phase and puzzler were not separated, as we will estimate the generalization error of the model if it were used to predict the results of the experiment repeated on a different day, possibly with different tangram puzzles, but not with a different setup (i.e. same phases and roles). See discussion for further comments on this.

2.3 Evaluation

To evaluate the models we used the MSE (Mean Squared Error) since our target variable is ratio, and we are interested in getting as close to the "Frustrated" values as possible, but the exact value isn't important. We are using MSE instead of e.g. mean absolute error, as the interpretability of the errors isn't too important for our case, where we are more interested in p-values. We also do want to punish bigger errors more than smaller errors. For clarification, MSE was also used to evaluate logistic regression.

2.4 Statistical tests

To test if there are significant differences between the different models, we use a Friedman test, as the errors of each model aren't normally distributed, and therefore we cannot use repeated measures ANOVA. We also tested the normality of the logarithm of the data, and the square root of the data, but it did not help. The repeated measures version is used as the errors of each model come from testing the same datapoints. Afterwards, we did post-hoc comparisons, using both setup II as described in Tue Herlau [2023] and a normal paired t-test. Note that the error means can be assumed to be normally distributed cf. the central limit theorem, as we have 168 errors for each model. Neither the Friedman nor the paired t-test account for variability in the training data, meaning they evaluate the difference between models **on this specific training data**. Setup II does account for variability in training data but has very low power, which is the reasoning behind doing both.

For all our p-values, our Bonferroni-corrected significance level will be 0.0016, as we calculate 31 p-values in total. This totals to a 0.049% risk of type I error in this entire assignment.

[1]The repository can be accessed directly at this link: `https://github.com/LucasLista/02445---Task-2`

2.5 Model robustness and consistency

To quantify the robustness to changes in data of the 6 models, We evaluate the standard deviation over the 8 folds' mean validation error. A high standard deviation indicates that the model performance depends a lot on the specific data (low robustness), and a low standard deviation the opposite. Obviously, this metric is also affected by the variance of the data itself, so it is solely meant to be compared between models.

To quantify the consistency of the models, we calculate the standard deviation of the test error when training the same model on the same fold 10 times. A high standard deviation indicates low consistency, and a low standard deviation the opposite.

Note that these values are only calculated to give the reader an overview of the models' performances, and will not be used to conclude anything. This is due to the fact that none of the models outperformed the baseline on average, and thus the variability of their performance is less interesting. For this reason, CIs weren't calculated, as that would either raise the risk of type I error which could mislead the reader, or lower the Bonferroni-corrected significance level intruding on our other tests.

3 Results

3.1 Model performance metrics

	Baseline	Ridge	KNN	RF	ANN	Logistic
Mean Error	4.478	20.191	4.431	4.821	7.132	6.554
Consistency Std	0	0.049	0.125	0.413	1.727	0.184
Robustness Std	1.725	38.022	1.548	2.037	1.877	2.814

Table 2: Consistency is evaluated by training the models on the same fold 10 times, and calculating the standard deviation of the resulting validation error. Robustness to changes in data is evaluated by the standard deviation of the mean error over the 8 folds. Note that CIs are not calculated, so we do not know whether the differences are signficant. See table 3 and 4 for difference in means statistical tests.

The consistency metrics of our models follow our intuition, the more random the model, the less consistent. The least consistent model was the ANN, which makes sense with random initialization and stochastic gradient descent.

The robustness holds more surprises, with the linear model being least robust by a substantial margin. Looking at the individual errors, it seems the linear model performed very poorly on especially fold 3. One could imagine this was due to the fact that the training data only held datapoints within a certain region of the input space, and when the model was tested outside this region, it extrapolated wrongly, which caused very high errors. The ANN was also in danger of this, but may due to its complexity and short training time have learned to use the input features sparingly. The other models are not in danger of this as they are bounded by the actual frustration levels reported by participants.

3.2 Friedman test

The Friedman test resulted in a test statistic of 25.153 and a p-value 0.00013. This is statistically significant, meaning that at least one of our models does not have the same performance as the others.

3.3 Setup II post-hoc comparisons

See table 3

3.4 Paired t-test post-hoc comparisons

See table 4

	Baseline	Ridge	KNN	RF	ANN	Logistic
Baseline	-	Mean=-13.75 p=0.518	Mean=0.05 p=0.826	Mean=-0.37 p=0.257	Mean=-2.53 p=0.091	Mean=-1.88 p=0.167
Ridge	Mean=13.75 p=0.518	-	Mean=13.80 p=0.513	Mean=13.38 p=0.531	Mean=11.21 p=0.578	Mean=11.87 p=0.568
KNN	Mean=-0.05 p=0.826	Mean=-13.80 p=0.513	-	Mean=-0.42 p=0.368	Mean=-2.59 p=0.067	Mean=-1.93 p=0.171
RF	Mean=0.37 p=0.257	Mean=-13.38 p=0.531	Mean=0.42 p=0.368	-	Mean=-2.16 p=0.159	Mean=-1.51 p=0.278
ANN	Mean=2.53 p=0.091	Mean=-11.21 p=0.578	Mean=2.59 p=0.067	Mean=2.16 p=0.159	-	Mean=0.65 p=0.596
Logistic	Mean=1.88 p=0.167	Mean=-11.87 p=0.568	Mean=1.93 p=0.171	Mean=1.51 p=0.278	Mean=-0.65 p=0.596	-

Table 3: Matrix over setup II post-hoc comparisons. The means indicate the row model - coloumn model error. Red indicates that coloumn model is better, green that row model is better. Note that all of the p-values are insignificant.

	Baseline	Ridge	KNN	RF	ANN	Logistic
Baseline	-	Mean=-15.71 p=0.00057	Mean=0.05 p=0.80718	Mean=-0.34 p=0.22000	Mean=-2.65 p=0.00006	Mean=-2.08 p=0.00000
Ridge	Mean=15.71 p=0.00057	-	Mean=15.76 p=0.00043	Mean=15.37 p=0.00078	Mean=13.06 p=0.00262	Mean=13.64 p=0.00218
KNN	Mean=-0.05 p=0.80718	Mean=-15.76 p=0.00043	-	Mean=-0.39 p=0.17658	Mean=-2.70 p=0.00000	Mean=-2.12 p=0.00000
RF	Mean=0.34 p=0.22000	Mean=-15.37 p=0.00078	Mean=0.39 p=0.17658	-	Mean=-2.31 p=0.00038	Mean=-1.73 p=0.00053
ANN	Mean=2.65 p=0.00006	Mean=-13.06 p=0.00262	Mean=2.70 p=0.00000	Mean=2.31 p=0.00038	-	Mean=0.58 p=0.43154
Logistic	Mean=2.08 p=0.00000	Mean=-13.64 p=0.00218	Mean=2.12 p=0.00000	Mean=1.73 p=0.00053	Mean=-0.58 p=0.43154	-

Table 4: Matrix over standard paired t-test post-hoc comparisons. The means indicate the row model - coloumn model error. Red indicates that coloumn model is better, green that row model is better. A light color indicates insignificance, and a darker color indicates significance. The significance level used is 0.0016 (Bonferroni corrected).

4 Discussion

We did find a statistical significant difference between models trained on this training data. None of the models were significantly better than the baseline model though, with Ridge regression, ANN, and Logistic regression being significantly **worse** than the baseline model. Only KNN performed better than the baseline, but there's not enough data to confirm the effect, and it's rather small anyway. Considering the wide range of models used, and the hyperparameter optimization done, it indicates that the heart rate data does not hold useful information for predicting frustration level. Interestingly, logistic regression performed better on average than ridge regression, almost to a significant degree. This is unexpected considering that classification is rarely used in regression problems, but may be a side-effect of the models being worse than the baseline (and the logistic regression making more "safe" choices).

The p-values from setup II are all insignificant, meaning that we cannot confirm that the same conclusion would be made with different training data. The power of this test is very low, so it does not indicate that there isn't a difference either.

4.1 Generalizability

When we chose to separate cohorts, rounds and individuals, but not phases and roles, we based the decision on the idea that the models will be used for predicting results of further experiments, where the same phases and roles are used. The reasoning behind this is that the data is collected from a very specific scenario. Participants know they are in an experiment, and are told to perform very specific tasks. There is no guarantee that this is corresponds to a person's normal experience of frustration in everyday scenarios, that possibly aren't as intriguing or fun. Possibly, they would rate their feelings differently if they were randomly asked to fill out a questionnaire throughout the day. The generalization error will therefore most accurately represent the error of predicting responses in a future experiment, conducted in the same manner. This is an important possibility, as one might see better correlation between heart rate data and frustration level in other scenarios.

In the case that one wants to generalize this study to more scenarios, it would be sensible to separate phases. Specifically, separate phase 1 & 3 from phase 2, as participants are asked to rest in phase 1 & 3 and are tasked with the puzzle in phase 2. Therefore, if separated, the generalization error could represent the error of predicting the frustration level of a human in a "new" scenario. The same argument goes for the role (puzzler) of the participant. However, while it would be a better estimate, there is still no guarantee that the fact that people are participating in an experiment is having a larger effect, that makes the correlations between target and feature different from other scenarios. For example, people may change their "frustration scale" depending on whether they are expecting to be frustrated, or for any reason convince themselves that they should be frustrated. There could be many cognitive factors at play. Do note that with more separation of groups, the theoretical generalization error will only increase, as we are extrapolating more. We should therefore expect to see the models performing even worse than they are now.

4.2 Areas of further research

4.2.1 Input features & ANN-variants

While we tried many different types of models, it is possible that using less of the input features, or trying more ANN-variants could lead to better performance. It does seem unlikely based on the fact that our other models generally performed so much worse than the baseline, but it should be mentioned. Especially, one could try classification ANNs, to examine if the logistic regression performance was caused by chance, or because classification models have an advantage for this task.

4.2.2 Effect of using summary statistics

While this report seems to indicate that heart rate data does not provide any information that helps predict the experienced frustration level of the experiment participant, it may also be because we are using the summary statistics instead of the raw data. One could imagine that certain patterns in the heart rate data would indicate frustration, but wouldn't change the summary statistics in any discernible way. It does seem rather unlikely, and it would probably make the model even more specific to this experiment, but it should be mentioned. A convolutional neural network would be good at recognizing those patterns if there were any.

4.3 Conclusion

The heart rate summary statistics do not seem to hold any predictive power of the self-reported frustration of participants in new experiments of the form described in Sneha Das [2024]. However, we did not have enough data to significantly confirm this conclusion for new training data, and we cannot with certainty generalize our conclusion to other scenarios. The KNN model, as the only one out of our 5 models, seemed to hold potential for beating the baseline, but only by a negligible amount. We note the possibility of using the raw heart rate data, as well as testing less input features and more ANN-variants, to potentially provide the last necessary information for a valuable predictive model, but find it unlikely due to the consistently poor models.

References

Carlos Ramos Gonzalez Line Clemmensen Nicole Nadine Lønfeldt Sneha Das, Nicklas Leander Lund. Emopaircompete - physiological signals dataset for emotion and frustration assessment under team and competitive behaviours. *ICLR 2024 Workshop on Learning from Time Series For Health*, 2024.

Morten Mørup Tue Herlau, Mikkel N. Schmidt. Introduction to machine learning and data mining. *Lecture notes, Spring 2023, version 1.0*, pages 215–218, 2023.