Project Title: Predicting Fatigue From Performance

Team Members: Nastasia Klevak
Emails: nklevak@stanford.edu

1 Motivation

In this project, I am working to understand the mechanisms that lead to cognitive fatigue building in people over time, and whether we can predict them using machine learning methods. This is an application project which uses data from two cognitive psychology datasets. These datasets both come from the same behavioral experiment (conducted at two different times with slightly different parameters), which were designed to investigate how task switching and performance interact to influence future performance and cognitive fatigue. In each experiment, participants were taught to play 3 games (spatial recall, digit span, and a simple rest game). Then, participants had 30 epochs of 10 trials each. Every epoch they were told whether the next epoch would be a switch or a stay (switch meaning digit span if they were currently playing spatial recall, and vice versa). Additionally, in between every epoch there was a rest period where participants could rest from 1 to 20 trials (as long as they wanted) in exchange for points they lost from an initial bank.

In my analysis of these datasets prior to this class, I've primarily looked at the interaction between previous performance and task switching on next-epoch performance using generalized linear models. This has had interesting results but has not tackled an important remaining question: how does fatigue (as measured by the amount of time people choose to spend resting) fluctuate through the course of the experiment, and what factors play a role in these fluctuations?

To answer this main question, I am using machine learning techniques from this class (and some extensions) to see how well we can predict a novel subject's rest length at every epoch of the experiment. Existing literature suggests that cognitive fatigue grows over time, and is at least partially a function of how well a person has performed on a task thus far. As a result, I'm interested in looking at how well different methods work to predict rest length, and which features are key to better predictions.

2 Methods

So far, I have run my main baselines and two extensions. For all of these runs, I've used two different methods of splitting the data. One is using the original dataset as a training set, and the replication dataset as a test set. I did this to get at the question of how related the two datasets are and if the same metrics are meaningful for prediction. The other split I used is taking 90 percent of the original dataset subjects and 90 percent of the replication dataset subjects, and using these combined subjects as the training set. The rest of the subjects (10 percent of each dataset) were used as the test set. I did this to see if training on data from a mix of both datasets was better for the baselines.

For my baselines I did Ridge Regression first. I used k-fold cross validation with 10 folds to tune the alpha parameter. I chose this method because it is a simple baseline to understand how well we can predict a subject's rest length in an epoch from simple features. Next, I did gradient boosted decision trees with the same training test splits. For this, I used k-fold cross val to understand how many trees were optimal to include, and at what depth. I chose this method as an alternative simple baseline for understanding how simple features can predict fatigue.

Moving on from baselines, I repeated the same methods but added in features that represent the history of the experiment (i.e. amount of rest in previous epoch, performance in last epoch, rt in last epoch). I'm doing this because it is believed that fatigue builds over time and depends on prior fatigue levels. As a result, I predicted that it would significantly boost my baseline performance to have a representation of how people were performing and how fatigued they were earlier in the experiment.

3 Preliminary Experiments and Results

For my preliminary experiments, I have run my initial baselines (Ridge Regression and Gradient Boosted Decision Trees) on my data using both aforementioned data splits. I have done this with and without the history features, and compared the results across all of these. A figure of early results is

below, but I will continue making some changes to the models (i.e. feature selection and scaling) as I continue working on this project.

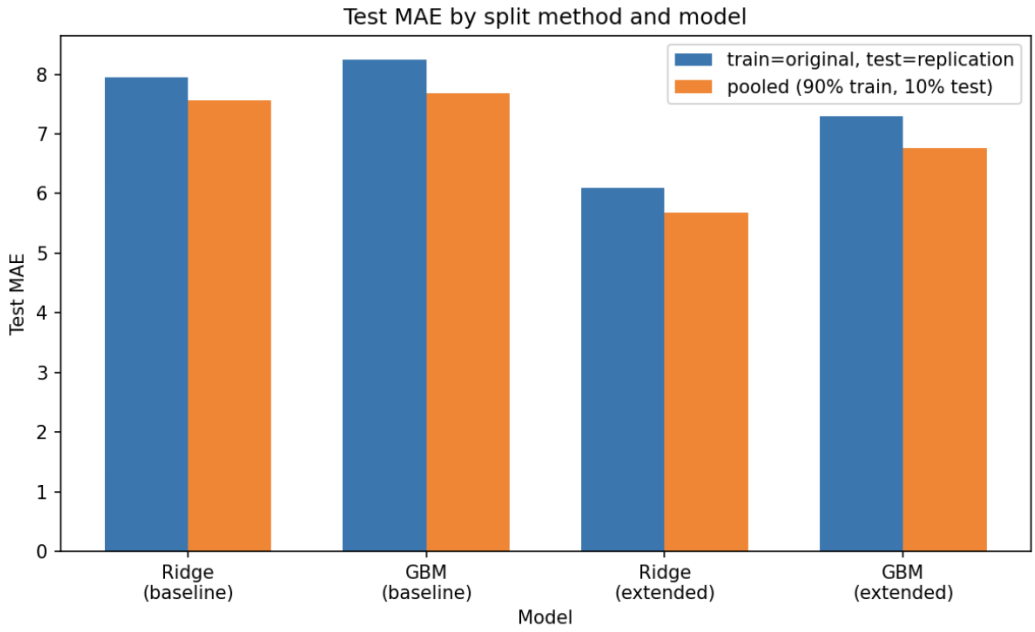

Figure 1: MAE (mean absolute error) comparison across baselines with and without history

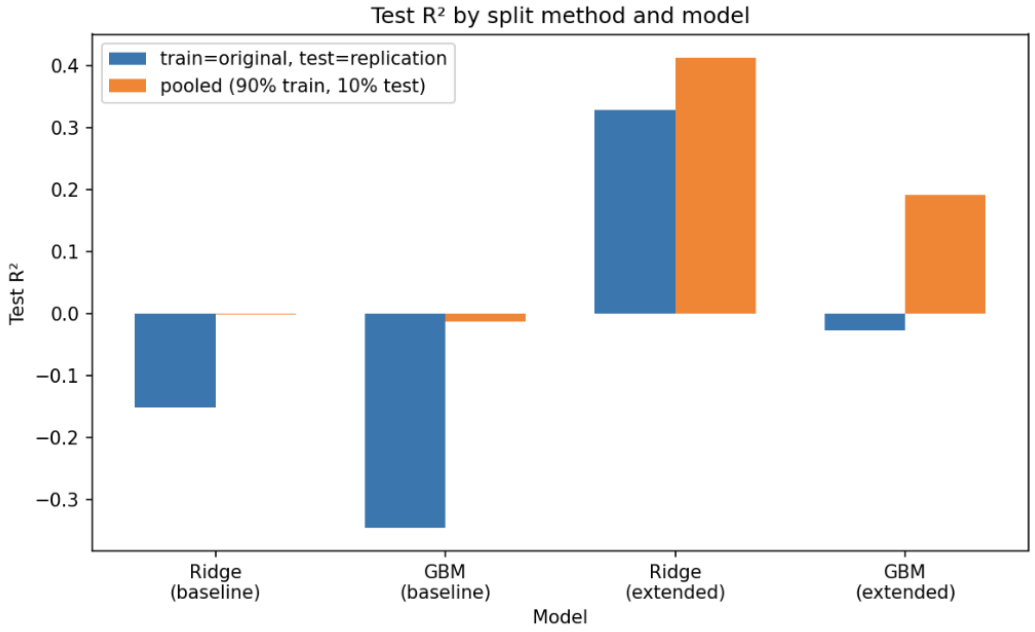

Figure 2: R squared comparison across baselines with and without history

So far however, as seen in the MAE figure, it seems as if the extended Ridge model (the one with historical features added) performed better (lower MAE) than the other methods. The same can be seen with the R squared values in the R squared figure. Additionally, it seems as if the pooled data split method generally performs better on the test set. This makes sense because there is data from both datasets in it (which have slightly different distributions), but is worth considering further.

4 Next Steps

Regarding the baselines, I plan to continue selecting the best features for those models as well as explore whether I need to scale the feature data in a different way. Additionally, I want to understand which features are most predictive for the test set, as well as calculate a predicted rest distribution for each subject and compare it to the true distributions in the test set for each baseline. Additionally, when looking at the MAE data more closely, it seems as if the predictions for the test set are always off by around 5-8 rest trials. It would be interesting to explore whether they are consistently over or under calculated with the different methods and try to understand why.

As a more complex predictive model, I'm interested in trying to fit a hierarchical LSTM in order to better incorporate all prior blocks when making a fatigue prediction. For this, I will train on entire experimental sequences for some subjects and try to predict all 30 rest lengths for other subjects. Because there might be too little data to properly train an LSTM, I am also considering simulating new data for the training set. Another option is to use more extended Ridge models, but included a lot more history (i.e. an array of time lagged performance metrics).

Finally, it would be interesting to cluster participants using k means clustering in order to understand different rest patterns, and to then compare these methods within each cluster to see if some are easier to predict than others.

5 Team Contributions

- **Nastasia Klevak**: I did everything since I am a single person team. I used the help of Cursor to figure out how to code up some of the baselines more efficiently. Also, some sections are copied from my earlier proposal (which I wrote and submitted to this class in the previous assignment).