**MIDS Capstone Project Development & Progress Survey**

**Menendez Joaquin**

**NOTE: Each member of the project should submit a separate survey. Please email your completed survey to Jana, Elchanan, Greg, and Ryan.**

The goal of this survey is to clarify the progress, goals, deliverables, and challenges of your capstone projects. It can be used as a diagnostic tool to identify the strengths and weaknesses of your projects, serve as a rough outline of your final write-up, and facilitate collaboration with technical and domain experts. It will also help us determine how best to support you between now and your final presentations.

The survey consists of a series of open-ended questions. Feel free to write as much as is necessary to convey your work and experiences to an interested non-specialist.The survey is split into five parts: (1) Goals, (2) Collaboration, (3) Data, (4) Modeling, (5) Deliverables.

**Goals**

**Q1: How would you currently describe the overarching goal of your project? What are you trying to understand, predict, model, etc.?**

In general terms, we are trying to predict or classify emotional states using Fitbit data, surveys, neuroimaging data and demographic data. We are still debating what would be the best approach to do so. Right now, we have agreed to predict the emotional valence (we don’t decide yet if we are going to construct a binary classification or a regression model).

**Q2: Projects evolve in time. How have your goals changed over the course of this project?**

I think that we had encounter some limitations with the data and that has made a little difficult to advance.

At the beginning we were considering the idea of developing an app (or ‘pseudo-app’). In this app the user was going to be able to have access to the prediction, and also be ‘notified’ when the algorithm was predicting that he was stressed.

Right now, we are not working anymore with stress. We switched to emotional valence, with a focus on negative valence. Also, we are not sure if we are going to be able to make this app.

**Collaboration**

**Q1: How often do you meet with your fellow team members? How often do you contact your non-academic partner?**

Once a week on Monday. We interact also on Friday once every two weeks. We talk via slack 2-3 times the week.

**Q2: Describe the division of duties between yourself and your fellow team members, and yourself and the non-academic partner. What do you expect from others, and what do they expect from you?**

Sicong has been working on the model and on feature engineering.

Mikella helped me to scrap data from one of the experiments. Currently she is working on the brain data and is our liaison with Greg.

I have been working with data cleaning, formatting and integration of the different sources of data.

**Q3: Are you happy with your collaborations thus far? If improvement is needed, what changes do you think are most important?**

I am happy with the team.

I think we started in a very disorganized way, now we are trying to correct that.

I would like to have some guidance regarding the theory background of the project, in other words if what we are doing has sense. I feel that nobody really works on emotions, so we are working using lay terms.

**Data**

**Q1: The cornerstone of Data Science is the data. Describe carefully the data you started with. What format did it come in? Were you expected to collect it, or was it given to you?**

The data was provided to us and it was a mess. It did not have proper documentation, had different formats, scales (in the case of the surveys) and inconsistency in how certain things have been recorded.

The raw data provided was in a .csv, .xls and .xlsx format

The data is composed by two datasets from different experiments (DNN and R00), both recollected by researchers from the Motivated Cognition and Aging Brain Lab (Mikella's Laboratory).  
There are 226 different subjects in total (122 subjects in one (DNN) and 104 (R00))

Every dataset is a collection of four different types of data:

* Experience sampling data
* Survey data
* Fitbit data (activity and sleep)
* Neuroimage data (not available yet)

**Q2: In what ways did you analyze/clean/pre-process the data before attempting to apply analysis techniques? Was the data clean or messy? Did you discover anything about the data that was not obvious to you at first?**

As I mentioned about the data was messy.

We performed several EDA at the beginning of the project.

Interesting findings

* The activity and HR data highly correlates
* Some subjects present several missing days of activity/sleeping data.

**Q3: If you are still trying to process the data into a more useful form (feature selection, removing noise, etc.), describe the challenges you are facing.**

I have partially answered this on Q1

* Right now, we need to figure it out what criteria the researcher used to classify sleeping data. This is problematic because Mikella do not know this, and the people that took the data is no more in her lab

**Modeling**

**Q1: What kinds of models have you applied to your prediction/analysis task? Why did you choose those models?**

**From Sicong’s survey:**

1. Logistic Regression & Linear Regression: Baseline

2. Random Forest: Low variance error.

3. CatBoost: Deals with categorical data, perform better than other tree based boosting model (XGBoost, LightGBM).

I would like to try a Hierarchical Model as a baseline. This would allow us t understand what variables have more influence on the outcome and it will be a good baseline to compare against a opaquer ML method.

**Q2: Did your models give you the results you were looking for? Why do you think this was the case?**

Currently the best model gives specificity (detection of unhappy subject) to be 58% with survey data + Fitbit data, 56% with only Fitbit data.

However, if we guess at random, the accuracy would be 10.8%.

Therefore, the model enables us to detect if Fitbit users are in negative mood with accuracy over 4 times higher than random guessing.

For this model we did not used a lot of survey, demographic nor brain data. I consider that we could improve these results after solving the issues aforementioned.

**Q3: What conclusions/results have you arrived at, so far? What do you still need to determine in order to wrap up the project?**

There are no strong conclusions yet, we need to finish solving some data cleaning/integration problems in order to be sure what we are doing.

I would like to have a better theoretical background of emotions in order to be able to generate better insights.

**Q4: If you are looking to try new techniques in the future, what are they? What help might you need in implementing them?**

Hierarchical model – taking each subject as a group, within group observations are very correlated when compared to between group correlation.

LSTM – Despite being opaque it could be a good alternative to use the time-series data we have.

**Deliverables**

**Q1: Sponsors want relevant data science insights. Have you discussed the question of deliverables with your sponsor? What do they hope to receive from you?**

On our last meeting I explicitly asked this to Greg, in summary he told us that the simple fact of proving that you can predict negative emotions using Fitbit + demographics + some short surveys is enough. I think that the fact of not having a clear goal it’s also making things harder.

**Q2: What do you currently plan on delivering?**

A trained model that predicts emotional valence (it could be binary or a regression model).

I would like to offer an interpretable model (in order to provide some insights to researchers). Nevertheless, I also would like to offer a very accurate model, despites that this could imply a not interpretable model.

**Q3: How do you propose to “sell” your deliverable to the sponsor? Why should they be interested in your work? How does it compare to existing research/methods?**

Given the literature we reviewed, there are not models that predict negative valence (most of the model aims to predict stress). Having a model that predicts negative valence could be useful features for a wearable device company. They could use this data to develop a ‘mental health’ feature. It could be useful for cases of depression. In these cases, monitoring the amount of behavior could be useful (mostly for a therapist that works with a behavioral activation perspective), if we add a graph/visualization with negative emotions during the day, the therapist and the patient could develop a more personalized treatment.

**Q4: What deliverables are you still working on? What needs to be accomplished in order to be able to deliver them?**

We are still working on trying to get a model with a good accuracy level. Solving the data cleaning issues is fundamental for this.

Add neuroimaging data to our model. Our idea is to have two models, one with all the possible features and other one, simpler with less feature but most probable to deploy in real life (use Ftbit data, some demographics data, some short surveys data, etc). We need to wait for Mikella to finishing this analysis.

**Is there anything else we should know in order to help make your Capstone project as successful as possible?**

I think that we need more guidance. At the beginning we were working by us now with no explicit direction. I think that now we have things clearer but having someone that help us to do this could be useful. Maybe some one that ask us for specifics analysis, reports, etc.

We consulted with Michael and Kyle. They were useful.