1. What is the problem you are attempting to solve?

TalkingData, China’s largest third-party mobile data platform wanted to use their data to help its clients better understand and interact with their individual respective audiences. My goal is to create a model which can accurately predict users’ demographic characteristics based on their app usage, geolocation, and mobile device properties.

2. How is your solution valuable?

Being able to correctly identify the characteristics of one’s audience can save financial resources since a company can focus on a certain demographic. Furthermore, advertising can become more personalized and perhaps knowing the user’s details can help with building a better recommendation engine whether in ads or some other aspect like Netflix show recommendations as an example.

3. What is your data source and how will you access it?

My data source will be from this website:

https://www.kaggle.com/c/talkingdata-mobile-user-demographics

There are multiple csv files which I will put to use.

4. What techniques from the course do you anticipate using?

I will use Pandas to clean the data set, removing blanks and NaNs. Afterwards, I will check to see if there are any class imbalances that I will have to address. I will use some exploratory data visualization in order to see if the distribution of data is normal or not and what that could potentially mean.

I will also have to generate some more features by turning categorical data into a usable numerical output. Depending on the outcome of my original data set I may have to generate more or less features.

Afterwards, I will use K-means as another way to explore the data and tell me what the most important components in terms of variance explanation are.

I will use the supervised learning techniques that I have learned in the boot camp after I create a train/test split with 25% holdout and search for the best parameters using GridSearchCV:

- Naïve Bayes Classifier, KNN Classifier, Decision Tree, Random Forest, Logisitic Regression (w/ L1 or L2 Regularization), Gradient Boosting Classifier, Neural Nets(MLP, RNN, CNN).

If my results aren't up to par, I may use PCA or SelectKBest as feature selection methods to see if they could help increase the accuracy of my models.

I will evaluate my models accuracy with a 5 fold cross validation, a confusion matrix, and a classification report. All models will be evaluated and critiqued in terms of computational time, and accuracy.

5. What do you anticipate to be the biggest challenge you’ll face?

Computational time will definitely be the biggest challenge I face. Hopefully I can remedy this by using the Google Cloud computing service. Perhaps the greatest challenge will be facing any unknown complications during this process where I can’t immediately think of a solution.