**HUDK 5053 Feature Engineering Studio**

**- Deliverable 5: Method Description (November 5, 2016) –**

M.S. Learning Analytics

Sunmin Lee

*Research question: Build a prediction model to find out key features of drop-out students in K-12.*

**1. Method description**

In the deliverable 4, I was able to make a simple visualization to identify the distribution of extracted data that I decided to use (i.e. Educational Longitudinal Survey 2002). This week’s mission is to plan the methodologies that I can use for this research question on the basis of consultation with Dr. Lang and classmates. Since my data contains only binary data as the variables represent the various reasons for drop-out students, I had many limitations in choosing different methodologies.

Upon discussion with Dr. Lang, I decided to use the Principal Component Analysis (PCA) to reduce the dimension of variables, which currently consist of 21 features. Although it is easier to run PCA in SPSS statistical package, I decided to use R Studio to practice the new skills. Initial data cleaning is done with M.S. Excel to run faster in R Studio. After running PCA, I am expecting to identify how many features are essential in the end.

Once I figure out the key number of features, I will run the prediction model such as regression and/or decision trees to measure the best combination of variables that predicts the outcome.

**2. Challenges discussed with peers**

One of the biggest challenges that I am confronting now is that I didn’t decide what to use for outcome (predicted) variables. Although I will be able to identify key predictors, I am not sure what my ultimate goal is in using prediction model. This solution needs more exploration in the dataset, determining which variable to use for predicted outcomes.

Another problem that I need to solve is that I might need to merge different datasets provided within ELS 2002. For the time being, I only extracted questions (i.e. variables) relevant to drop-out topics. However, if I want to choose a certain variable for outcome results, I might need to combine other datasets, which I am not sure if we have a common standard (e.g. student id) that I can use. More exploration of datasets is required.