Elliott Wooton

Project: Predict Student Performance from Gameplay

1. **Intro:**

Of the many, currently on-going challenges hosted by Kaggle, the dataset used in this project will be taken from The Learning Agency Lab. This dataset is a compilation of every action a student does in a game while answering questions. Using this information, a model will be created that can be fed actions a student has done throughout a game session and be able to guess whether that student passed or failed. This white paper is intended to provide an overview of the data per-processing and analysis, as well as the modeling and evaluation of a machine learning algorithm.

2. **Dataset:**

A test and train dataset is available from The Learning Agency Lab hosted on Kaggle, under the Predict Student Performance from Game Play challenge. The dataset comes in the form of a csv with 20 explanatory columns and a response (whether the correct answer was chosen or not.)

3. **Data Preparation:**

The data comes with 20 columns and the target variable, with many columns only containing things such as session ID or row index. After cleaning out unnecessary data, the data must be check for noise such as outliers and repeat information. Given the nature of this type of dataset, there is likely to be little in the way of repeat information, however outliers could certainly occur in categories such as elapsed time.

4. **Exploratory Analysis:**

EDA is very important in data mining in order to gain understanding of the data to aid in algorithm selection. There is also the ever important problem of Multicollinearity, which correlation matrices can help discern. Pairwise scatter-plots and summary statistics will give shape to the data, showing if any normalization is needed to be done for parametric machine learning methods.

5. **Modeling:**

Given that the goal of this project is binary classification, algorithms have been chosen based on that criteria. In particular, Random Forest Decision Trees, Logistic Regression and Ada Boosting have been chosen. The other main factor regarding the choice behind the algorithms is the large number of rows and columns. 20 explanatory variables is quite a large amount that will more than likely need to be scaled down through PCA or regression. There are also 13.2 million rows in the training dataset, meaning the algorithm chosen will need to be significantly light weight in order to run in a reasonable time frame.

6. **Evaluation:**

Considering this model is going to be a binary classifier, evaluation metrics are going to include: Accuracy, Recall, Precision and F1 score. The evaluation will be chosen based on findings from the explanatory data analysis step.