Nadeem Nasimi
Michael Hammer

# No8am Datamining Final Project - Phase II

## Accomplishments

### Creating Plots

We created a variety of bar plots to help us better understand the distribution of our data. First, we looked at the relative popularity of courses chosen in each department. We did this by plotting each department versus the number of times a course was chosen in each department. Then, we investigated the popularity of courses within a department, for several departments at Bucknell, by plotting each course in a department versus the number of times a course was chosen. Subsequently, we created plots that depict the number of courses and the number of sections that students put in their schedule. These plots gave us insight to the workloads that people had. Finally, we made a plot that showed the number of schedules made in each month. This plot allowed us to visualize when students were making schedules and thinking about what classes to take.

### Useful Datasets

Initially, we created a master dataset that held all of the information read in from our input file. This dataset was cleaned and held all the data that we would potentially use. From this, we created all of the other datasets. First, we created datasets for our plots. This included sets with the number of courses in a schedule, the times a schedule was created and more. Subsequently, we had a dataset that used binary to show if a schedule had a course. Finally, we made a dataset to show the departments that each schedule has.

### Association Rules

We were able to use the Apriori algorithm to create association rules from two sets of data: individual courses and departments. The first dataset showed each course in a student's schedule. The association rules created from it showed which courses were strongly related. The second set of data included the departments that a student had in their schedule. It's association rules showed which departments were strongly related.

### Predictive Models

We used several models to predict whether a department should be in any given schedule. The models we have used so far include Naive Bayes, Random Forests, Decision Trees, and Neural Nets. We had excellent results on several models when predicting if CSCI should be in a schedule. We plotted the performance of these models and evaluated the performance of each one.

## Challenges

### Refactoring

As our code base has increased, the need to refactor our code has become increasingly important. We broke up large code chunks into smaller ones to allow for better documentation. Additionally, we grouped all data frames, and other objects to be passed into functions, into code chunks at the top of the R Markdown file so the data creation process is separate from our results and analysis. This made our analysis clear and concise. Another step we took to refactor was to move repeated code into functions. This gave us a chance add modularity and reduce the length of our sections. Also, moving code into functions allowed us to reevaluate the algorithms we wrote and reduce their complexity.

### Creating Useful Datasets

We found difficulty in creating useful datasets. First, we had to create several datasets and we did not initially know all the information that we would need. While we developed, we would reuse previously created datasets or create new ones with the information that we needed. Second, it was difficult to alter datasets so that certain R functions could use it. For example, for each dataset the predictive models needed us to factorize each column, have multiple levels, and make sure that the column we were predicting could was not a number. It was difficult to discover this and convert our datasets into that form. Finally, it was difficult to organize the datasets. At this point we had created many datasets so we had to refactor and organize our code in order to keep track of them.

### Understanding R Functions

As we were experimenting with the functions for Apriori and predictive models, we ran into many error messages that were often difficult to decipher. In order to overcome this challenge, we read the help pages for these functions and researched online to find how others overcame these obstacles. Most of our issues were resolved by transforming our data into a format that was accepted by the function we were trying to use.

## Future Work

As we approach the final stage, we plan on improving our documentation in the R Markdown file to describe in detail what we are doing at each stage of the process and justify our methods. Additionally, we plan on tuning the input variables passed into our models to produce even better results.