Final Project

Luke Haws, Jared Cordova, Dario Fumarola, Ryan Messick

We have decided on a dataset from the University, but as we are still waiting to hear back about this dataset, we also have prepared a backup dataset we found online. Our primary dataset will (hopefully) be a collection of all alumni who have donated to the University in the last however many years back we can get our hands on. We visited the department Friday and were told to talk to Heather Meixler about the project, but she wasn't in that day. So, we emailed her and are waiting to hear back. We're hoping to be able to take a dataset of alumni (without any direct identifying information) and use features such as their graduating year, major, current field of occupation, political affiliation, and others, as well as if they've donated money to the university and when, and create a machine learning model that can predict how likely an alum is to donate to the school based on those features. This model will depend on what features we have available from the school's dataset, though, and we cannot fully prepare to build this model until we see the dataset.

Our backup option involves a dataset from Portugal we found online. This dataset (https://archive.ics.uci.edu/ml/datasets/student%2Bperformance) contains information about secondary school students in two Portuguese schools and their grades throughout the year. This dataset has a great number of features, such as age, family size, parental education, study time, and more, and many of them will require some feature engineering to extract useable data. Ultimately, the goal with this project would be to be able to predict a student's grades based on information about them. Our problem statement is this: Predicting student performance in secondary school prior to the academic year would allow a school system to cater to individual students' needs more precisely. This would allow for more students to perform well in secondary school, which could lead to more students attending university and likely contributing more to the economy.

This model would be useful for being able to identify students that might need extra tutoring or special attention before their grades plummet. If this model could be applied to students worldwide, then many people and countries could benefit from a model like this. This paper (http://www3.dsi.uminho.pt/pcortez/student.pdf) used this dataset and discusses how it could be useful in predicting student performance. The paper used decision trees, random forest, SVMs, and neural networks to predict student performance. Ideally, we would try to use various models and compare the effectiveness of each. We'd check their results by using the models we've learned of that they've used, and also implement KNN, Naïve Bayes, linear regression, and logistic regression.