Research Subteam: Presentation 1

Maxwell Bach, Okubay Gebrelibanos, Yili Hui, Arjun Patel, Manley Roberts

Research Question

- To use various subsets of course data related to performance in a course (ISyE 6501, Introduction to Analytics Modeling), combined with various machine learning/statistical models, to predict final success in the course.
- Once these models are trained, we will attempt to make conclusions about **particular predictors** of **success** or **failure** in the course.
- The ultimate goal is to evaluate these models against one another with a range of methods, including naïve average, null hypothesis statistical testing, and ultimately, Bayesian Hierarchical Modeling.

Master Plan

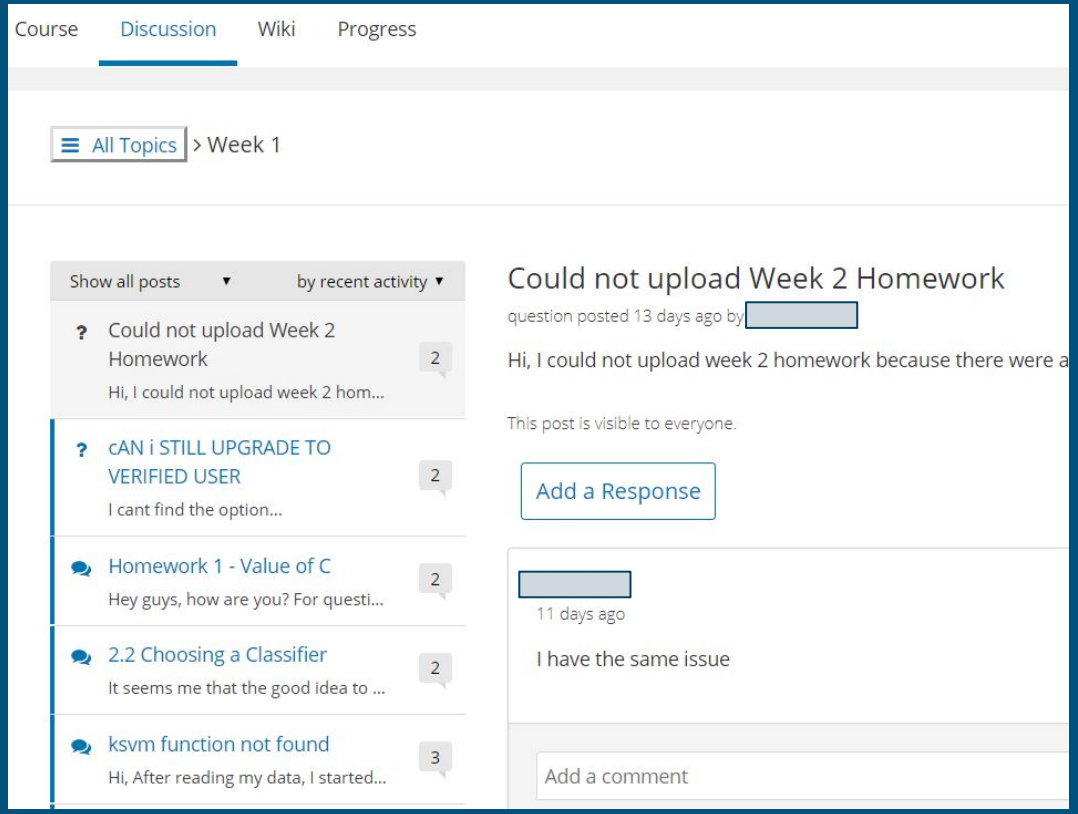
- Aggregate/clean data
- Create a library of models
- Evaluate model success

Phase 1: Data

- Clickstream Data: MongoDB
  - Percentage of Total Video Duration Watched
  - Forum Views
  - Active Days
  - Quiz Views
  - Exam Views
- Assignment Data: Edx/Canvas on PostgreSQL
  - Pre-submission Lead Time (how far before deadline submitted)
  - Total Quiz Score
  - Total Assignment Score
  - Total Number of Quiz/Homework Submissions
  - Change in Weekly Average Grade

Phase 1: Data (Continued)

- Forum Data (EdX) w/ Natural Language Processing
  - Number of Posts
  - Number of Replies
  - Average Post Length
  - Average Post Sentiment
  - # of Positive, Negative, Neutral Posts
  - Threads Started
  - Unique Words/Bigrams
  - Flesch Reading Ease
    - How easy to understand?
  - Net Votes Received



Phase 2: Models

- If we want to compare models head-to-head...
  - These models must attempt to answer the same question:
    - Course success
    - Success on a particular assignment or quiz
- We aim to train numerous machine learning/statistical models
  - We will group these models by the question they are answering
- We'll mix and match data feature sets with models
  - For example, **Adaboost** applied to **Forum Data** to predict **Overall Course Success**
  - Alternatively, **SVM** applied to **Clickstream Data** to predict **Quiz 2 Grade**

Phase 2: Models

- Possible Models
  - Classification Trees
  - Logistic Regression
  - Adaboost
  - SVM
  - Naïve Bayes
  - And beyond...
- Hyperparameter Tweaks
  - Modify the operation of models in a number of ways
  - Learning rates, number of iterations, style of cross-validation, etc

Phase 3: Evaluation

[# of Feature Sets] X [# of Models] X [# of Tweaks] = Many possible observations

- We want to figure out how *best* to predict success on a given portion of the course (or the entire course).
- Compare models with one another
  - Naïve Average
  - Null hypothesis significance testing
  - **Bayesian Hierarchical Method**
- Looking for the family of best models
  - Perform better than any other model

What we've done so far...

- Write research question
- Request data access - EdX, Canvas
- Establish common Docker build system
  - All models built for a generic Linux virtual image
  - Improves compatibility between developers' environments
- Split into sub-sub-teams
  - Clickstream: Arjun, Maxwell
  - Assignment: Manley, Yili, Okubay

What comes next

- Data Cleaning
  - Focusing down on the particular data we're looking for
    - Inputs
    - Outputs
  - Separating by Student Group
    - Audit
    - Micro-Masters
    - OMSA