Research Subteam: Final Presentation

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

Review: Master Plan

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

Final Data

- EdX SQL Server
  - Course ID (Restricted to ISYE 6501, Fall 2019)
  - User ID (Restricted to Verified)
  - Week Number (2 - 16)
  - Grade on Weekly Assignment
  - Grade Difference Between This Week and Prior Week
  - Time Remaining Upon Homework Submission
  - Passed Course
  - Final Grade

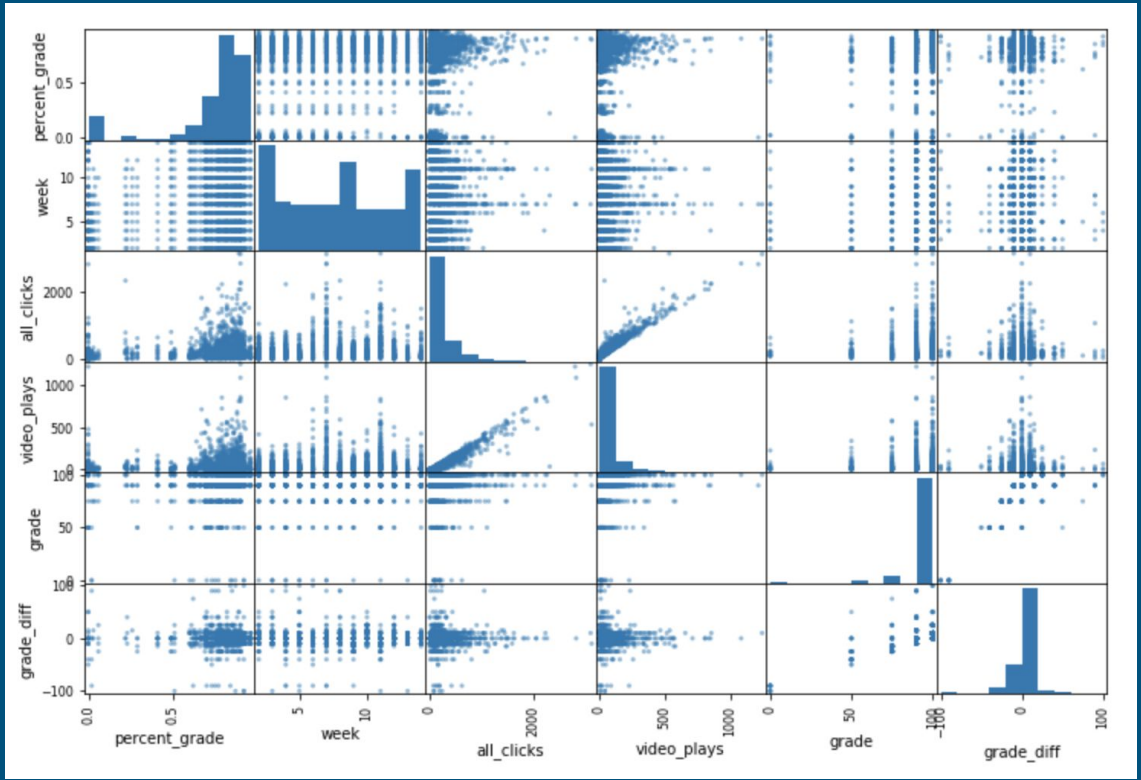
- EdX MongoDB ClickStream - Weekly Counts
  - seq_next (Navigate to Next Page)
  - seq_prev (Navigate to Prior Page)
  - seq_goto (Navigate to Specific Page)
  - play_video
  - pause_video
  - closed_captions_shown
  - problem_save (Save Interactive Problem Answer)
  - seek_video
  - link_clicked
  - Total Clicks

2043 distinct tuples

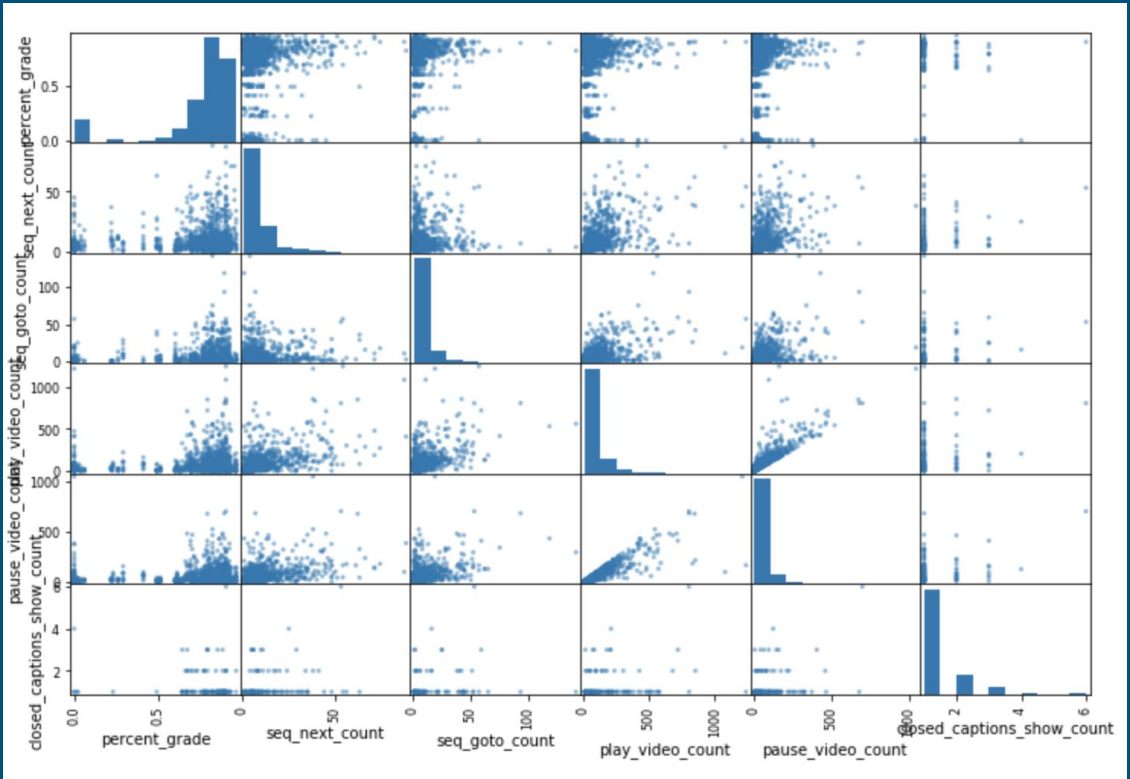
Some Basic Info

	all_clicks	video_plays	seq_next_count	seq_goto_count	play_video_count	pause_video_count	seek_video_count	problem_save_count	percent_grade
count	2043.000000	2043.000000	1534.000000	1114.000000	1639.000000	1630.000000	1432.000000	48.000000	2043.000000
mean	251.113558	61.966716	9.068449	8.191203	77.241001	50.800613	43.551676	9.208333	0.729227
std	313.500613	107.775512	10.379235	12.039553	115.325313	76.546371	109.116649	8.374120	0.248482
min	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	76.000000	5.000000	3.000000	2.000000	16.000000	10.000000	6.000000	2.000000	0.720000
50%	155.000000	26.000000	6.000000	4.000000	38.000000	24.000000	15.000000	6.500000	0.810000
75%	297.500000	72.000000	10.000000	9.000000	92.000000	62.000000	40.000000	17.000000	0.870000
max	3121.000000	1241.000000	88.000000	141.000000	1241.000000	1048.000000	1973.000000	27.000000	0.960000

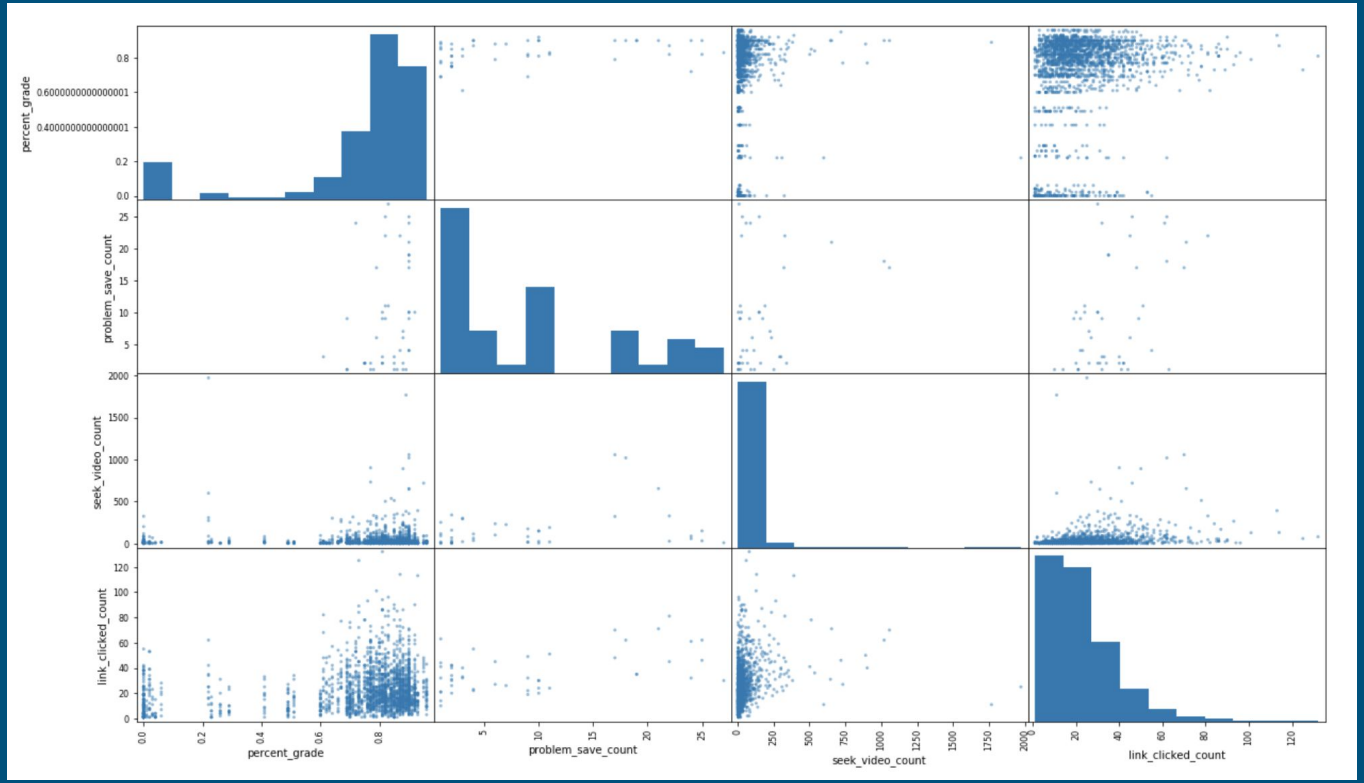
Correlation



Correlation



Correlation



Final Data - Xs and Ys

- EdX SQL Server
  - Course ID (Restricted to ISYE 6501, Fall 2019)
  - User ID (Restricted to Verified)
  - Week Number (2 - 16)
  - Grade on Weekly Assignment
  - Grade Difference Between This Week and Prior Week
  - Time Remaining Upon Homework Submission
  - Passed Course
  - Final Grade >= Median (0.81)
- EdX MongoDB ClickStream - Weekly Counts
  - seq_next (Navigate to Next Page)
  - seq_prev (Navigate to Prior Page)
  - seq_goto (Navigate to Specific Page)
  - play_video
  - pause_video
  - closed_captions_shown
  - problem_save (Save Interactive Problem Answer)
  - seek_video
  - link_clicked
  - Total Clicks

Key

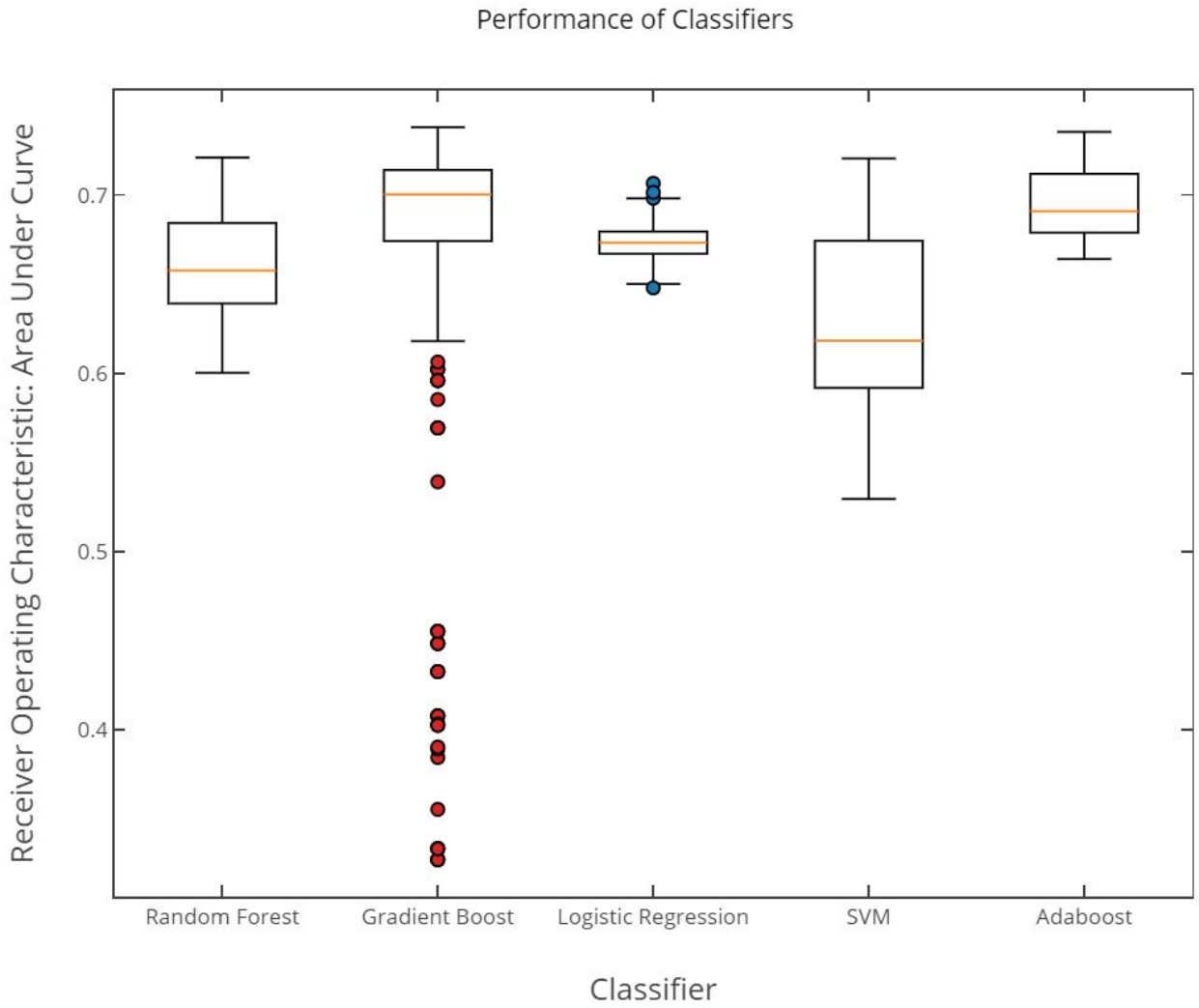
- X attribute
- Y attribute
- Unused Attribute

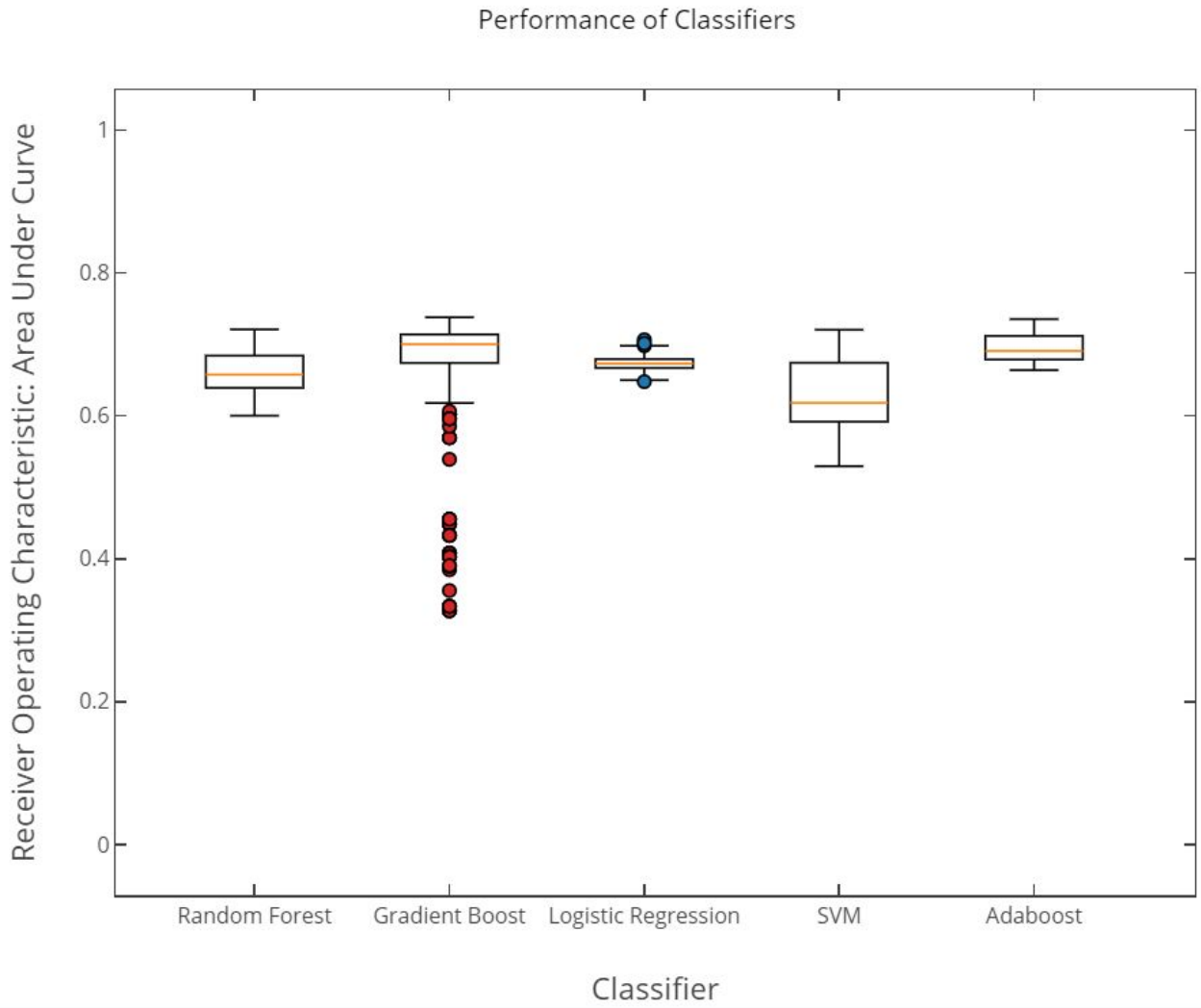
The Problem

- Given a week-long snapshot of a student's performance, can we predict whether or not they will do better or worse than the median score at the end of the course? (binary classification)
  - How do we predict it?
  - How well can we predict it?

Application of Models

- Random Forest
  - Imputer Strategy: Mean, Median
  - Number of Estimators: 5, 10, 100
  - Criterion: GINI, Entropy
  - Max Features: Auto, Sqrt, Log2
- Gradient Boost
  - Number of Estimators: 5, 10, 100
  - Learning Rate: 0.01, 0.1, 1, 10
  - Loss: Deviance, Exponential
- Logistic Regression
  - Penalty: l1, l2
  - C value: 0.01, 0.1, 1, 10, 100
- SVM
  - Kernel: Linear, Poly, Rbf, Sigmoid
  - C: 0.01, 0.1, 1, 10, 100
- ADABoost
  - Base classifier: Decision Tree Classifier, Depth = 1
  - Number of Estimators: 20, 50, 100, 200


Performance of Classifiers
Receiver Operating Characteristic: Area Under Curve
0.7
0.6
0.5
0.4
Random Forest
Gradient Boost
Logistic Regression
SVM
Adaboost
Classifier


Performance of Classifiers
Receiver Operating Characteristic: Area Under Curve
1
0.8
0.6
0.4
0.2
0
Random Forest
Gradient Boost
Logistic Regression
SVM
Adaboost
Classifier

Performance Summary

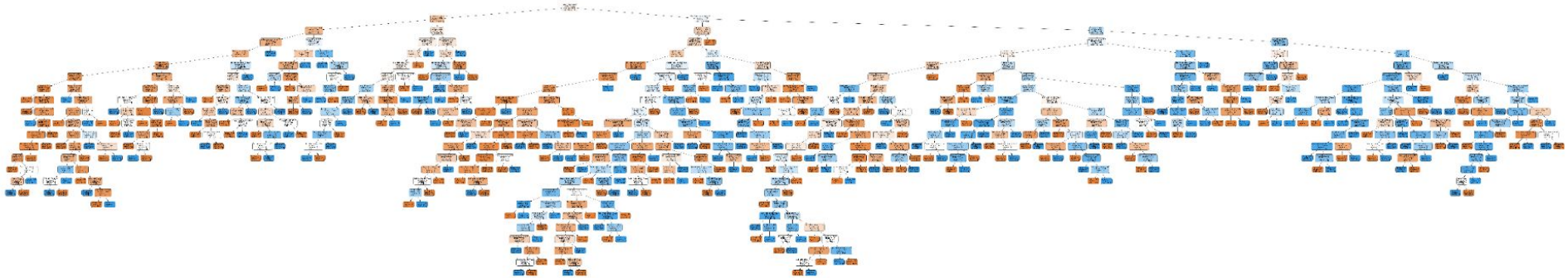
Model	Random Forest	Gradient Boost	Logistic Regression	SVM	ADABoost
Max	0.721029	0.73806	0.706591	0.72054	0.735447
Median	0.657685	0.700326	0.673301	0.618333	0.690892
Mean	0.659366	0.661896	0.6737	0.630556	0.695087

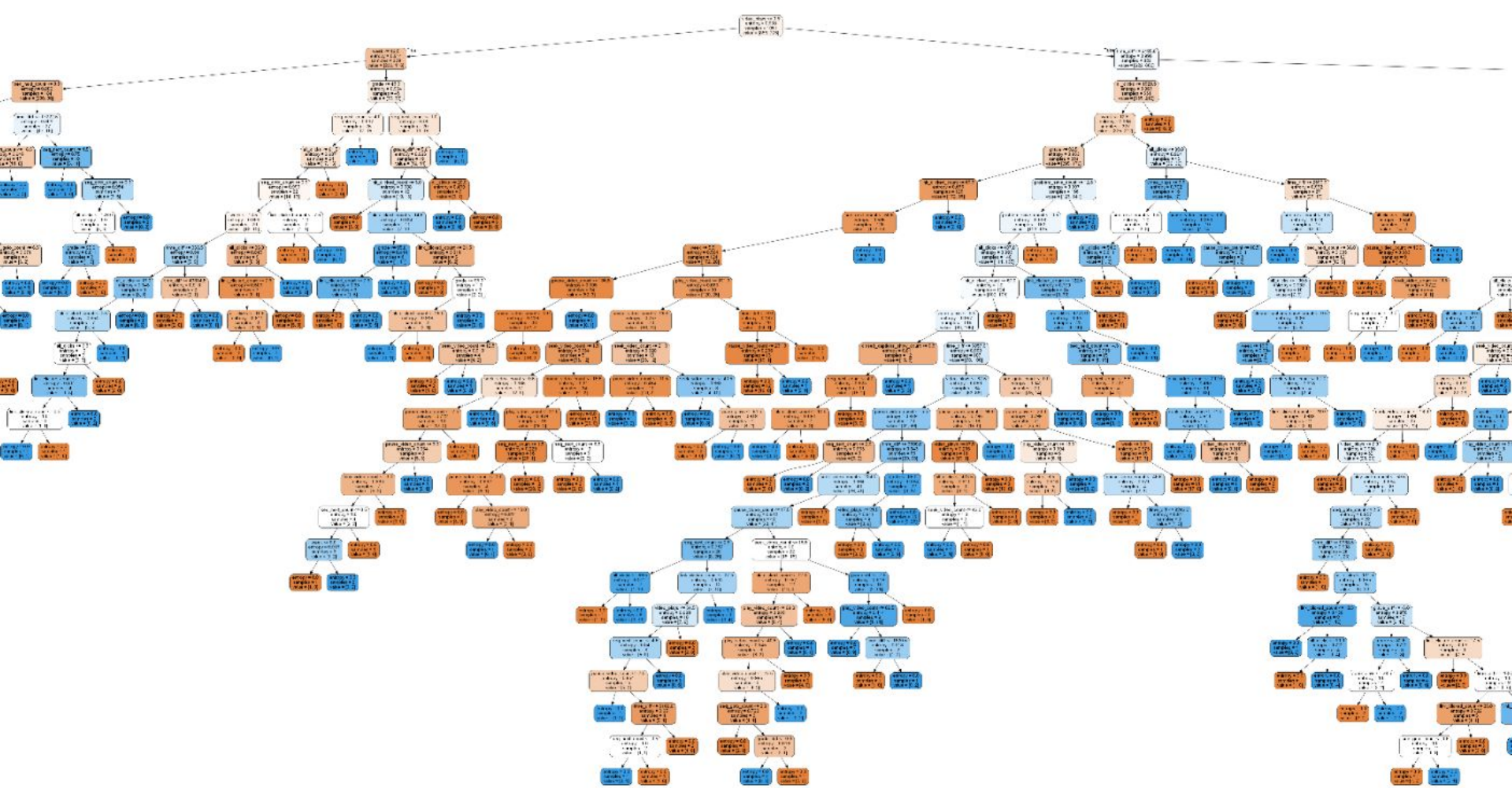
Optimal Model Parameters

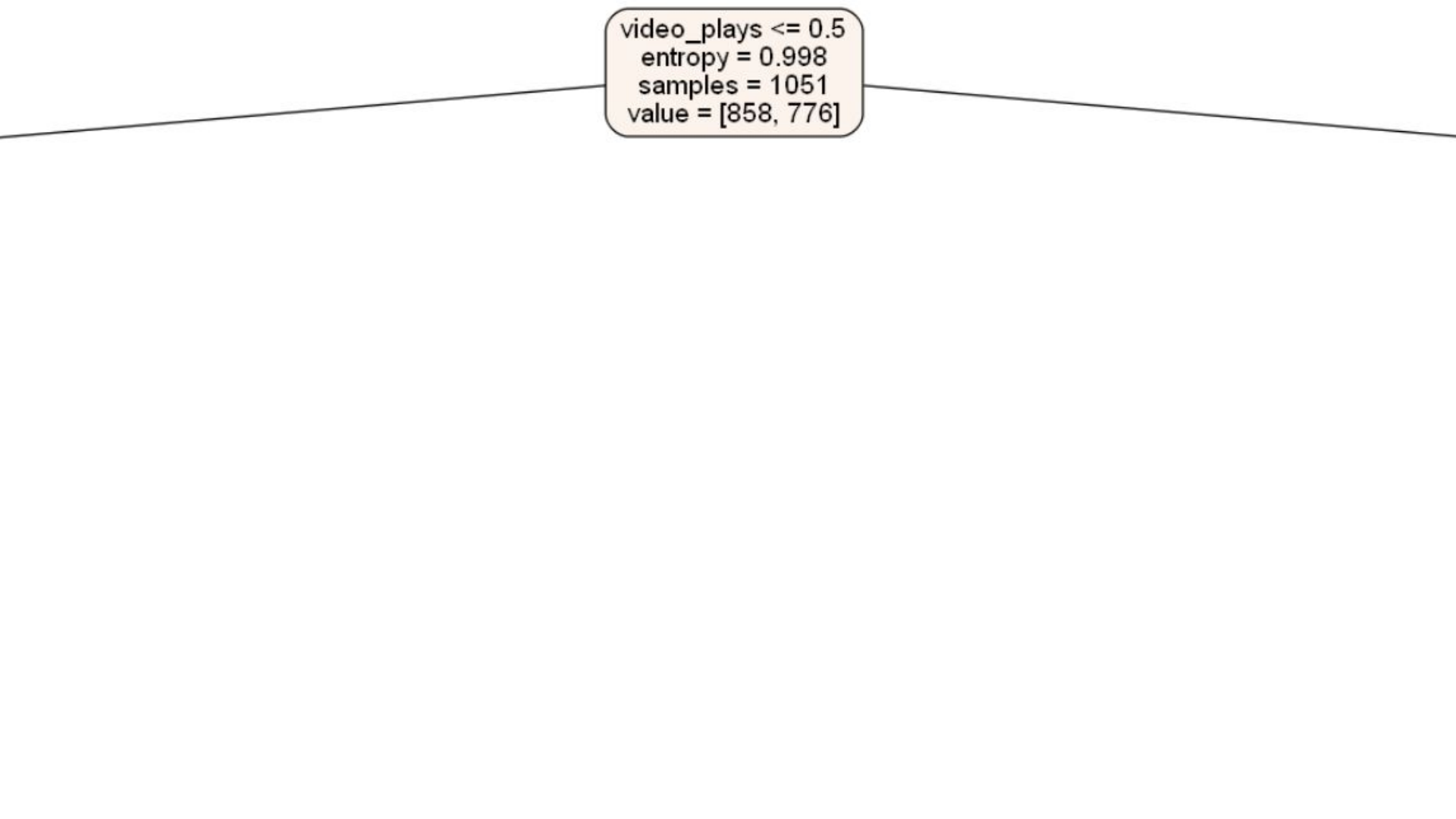
- Random Forest
  - Imputer Strategy: Mean
  - Number of Estimators: 100
  - Criterion: Entropy
  - Max Features: log2
- Gradient Boosting
  - Number of Estimators: 100
  - Learning Rate: 0.1
  - Loss: Exponential
- Logistic Regression
  - Penalty: L1
  - C: 0.01
- SVM
  - Kernel: Linear
  - C: 0.1
- ADABoost
  - Number of Estimators: 20

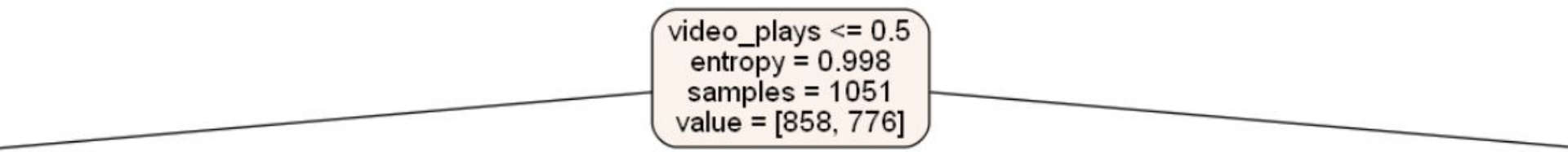
A Deeper Dive: Random Forests

1 of 100 Decision Trees, which each vote when attempting to classify data






video_plays <= 0.5
entropy = 0.998
samples = 1051
value = [858, 776]



Switching over to the actual image file...

What Comes Next?

- Bayesian Hierarchical Modeling
  - Implementation & Application
- Forum Data
  - This semester's EdX forum went primarily unused
  - Worth examining Piazza in courses that use it
- Data Subsets / Feature Selection
  - Not every feature is equally useful
  - It may be easier to obtain a smaller subset of features
- Predicting Success at Checkpoints
  - Use all student data up to 1 month, or halfway point, etc.
  - See at what moment we begin to get a good prediction
- Portability to Other Classes
  - CS 1301X
  - Boilerplate Code / Models