MACHINE LEARNING REGRESSION ANALYSIS OF EDX 2012-13 DATA FOR IDENTIFYING THE AUDITORS USE CASE

Mark Mueller and Greg Weber

Georgia Institute of Technology, Computer Science, Atlanta, GA

ABSTRACT

Predictive models are able to predict edX student grades with an accuracy error of 0.1 (10%, about one letter grade standard deviation), based on participation data. Student background variables are not useful for predicting grades. By using a combination of segmentation, random forest regression, linear transformation and application beyond the segmented data, it is possible to determine the population of the Auditors student use case, a population larger than those students completing courses with grades.

1. INTRODUCTION

Predictive regression models have been developed for prediction of student grades based on performance and other student data. The grade for a given class can be predicted with a standard deviation of 0.1, corresponding to approximately one letter grade. From this work, we determined important features for prediction which exclude student background information such as year of birth, gender and level of education. We find that an earned grade is dependent on participation: chapters, number of days participating, and other participation metrics.

Previous publications[1-8] describe a completion rate for MOOCs which is typically around 3% and always in single digits. It is suggested[1] that this 3% represents only one segment of learners, and that other learner use cases exist within the remaining 97%. Auditors may form a substantial portion of the committed, learning student population. In Duke University's paper[6], a plot of student participation suggests an auditor group which is twice as large as the traditional course-completing or "certified" group. The focus of this investigation is finding the auditor use case within the publicly available edX 2012-2013 dataset.[9]

1.1. EDX DATASET

The publicly-available edX dataset is a downloadable .csv file[9]. Provided data is "final", not subject to updates. This data consists of 640,000 rows, each representing a course taken by a student in 2012-13, and 20 columns containing student and course information. At the time, edX offered 16 courses created at Harvard and MIT. 100k rows of this data show an "inconsistent" flag. After removal of these inconsistent-flagged rows, the focus is on analysis of the remaining 540,000 instances.

Each row of edX data includes student information (year of birth, gender, level of education) and student ID is anonymized. Each row also contains course performance information such as nchapters (number of chapters accessed), nvideo (number of video actions taken), ndays (number of days accessing online material), start date and end date, and final grade. If a student achieved a high enough grade they are "Certified". Only about 3% students reach this status.
The header describes 4 types of students: registered, viewed, explored, and certified. The first

DOI :10.5121/ijite.2017.6301

category, registered, includes all students in the data. To be "viewed", a student had to log in just one time. The "explored" designation is given to students who view half or more chapters of course material. Finally, "certified" describes those students who achieved at least a minimum grade within the class (50% to 80% depending on the course). The "certified" segment forms the basis for further analysis.

Most of the previous analyses focus on the 3% of students who complete a course with a high grade which is usually considered passing in a regular classroom: >60%. However, two papers[1,6] provide more depth and insight into the other 97%.

1.2. PREVIOUS EDX ANALYSIS

Background literature includes analysis of student behaviors in MOOCs[1,3,5-7], general MOOC discussions[7,8,10-15], and analysis of the edX dataset or portions thereof.[2,4,5,9] None of these studies utilizes regression as a predictor of student grades, nor leverages segmentation for training and testing a model for application outside the segmented data. Thus, proposed methods applied to edX MOOC data here are novel.

Previous work describes MOOC course completion versus other use cases. The Duke Bioelectricity[6] paper shows data which may be interpreted to infer that the number of auditors in a MOOC is double the number of "certified" completers. The Duke paper, and others such as Greene[17], describe MOOC attrition curves plots. From these graphs and associated information, it is possible to compute the number of auditors within the Bioelectricity class.

The Bioelectricity paper does not address the topic of auditors, but by applying Kizilcec[1], who defines auditors as students who watch all videos but do not perform graded tasks, and by subtracting the number of course completers (313, Bioelectricity, Fig. 3) from those watching videos through the last week (1000, Bioelectricity Fig. 2), we can observe 700 auditors. Thus, the ratio of auditors to completers is about 2:1. The Bioelectricity relevant plot and bar graph are shown in Figure 1.

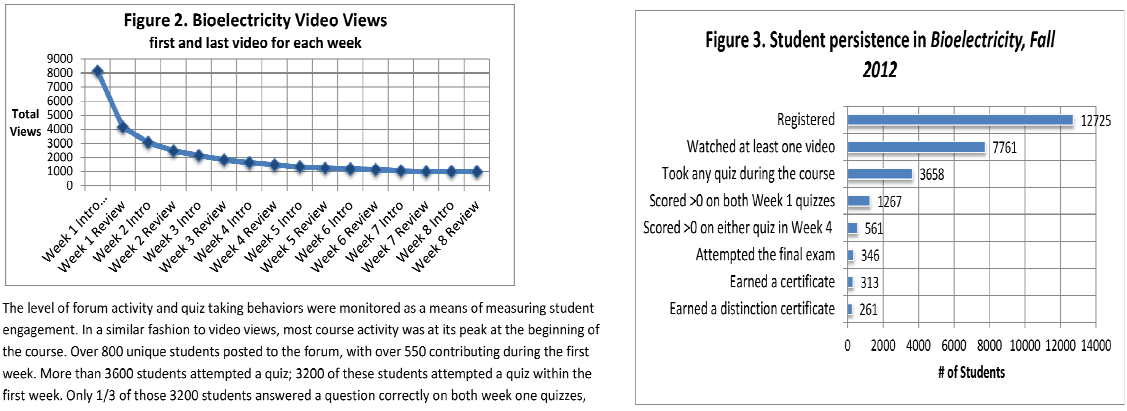


Figure 1: A plot and a bar graph from Bioelectricity[6] provide important clues toward understanding auditing. In the last week, there were ~1000 video views, while ~1/3 of viewers attempted the final exam and earned a certificate. The other 2/3 are auditors.

This discovered 2:1 ratio of auditors to "finishers" relies on a Kizilcec's definition of auditing as viewing videos through the last week of the course. We can apply a generalized version of this

definition to the edX data: auditors are students whose online participation is indistinguishable from that of the Certified group.

By segmenting Certified from non-certified data, it is possible to build a grade predictor. This predictor may be applied to the non-Certified data to identify auditors. Characteristics of auditors can be revealed by comparing histograms to those of the Certified group.[4,5]

2. PREDICTING COURSE GRADE USING MACHINE LEARNING REGRESSION ALGORITHMS

The data processing pipeline for predicting grade includes preprocessing, segmenting, training and testing, and applying the trained model to the non-certified population. Regression models were generated using Certified data. The training error and testing error are both less than 10%, or one letter grade (or 0.1 in the grade range from 0.0 to 1.0).

ML processing steps include

(1) segment the certified group.

(2) Create predictive regression models for predicting a student's grade (0.0 to 1.0) based on participation metrics and student background information.

(3) Fine tune model parameters for prevention of overfitting and for minimizing RMS error

(4) Apply regression models to the non-Certified data to determine the number of "auditors": students whose participation makes them indistinguishable (within the context of the regression models) from Certified students.

Models for predicting "certified" students' grades have been trained and tested. Best test results are about 10% (0.1) RMS error on the test set, or about 1 letter grade. Linear regression is the simplest model with the highest RMS error. Decision tree test error is about 0.12 while random forest and gradient boosted models' testing errors are about 0.1.

2.1. PREPROCESSING

Preprocessing of data includes: reading the .csv file, deletion of zero information columns, removal of instances whose "incomplete" flag is 1, conversion of strings into numeric values, replacement of NaN (not a number) values with appropriate default values, and normalization of nchapters. Because the number of chapters varies drastically between courses offered, this normalization creates improved models. Finally, the dataframe is converted to numpy arrays X, y for machine learning. For each row i representing one course taken, y_i is the grade for the course taken, while X_i is the vector of numeric variables for that course.

2.2. SEGMENTATION OF TRAINING/TESTING DATA

Certified instances comprise about 3% of the courses taken. Certified completion maps into the "bricks and mortar" traditional education use case of students who have participated and passed a course. By segmenting the Certified instances from others, we can build regression-based models for predicting students' grades y from other information X.
Without segmentation, distributions of participation and grade variables are centered near zero. In contrast, Certified course data maps into grades expected of traditional education, so that edX data can be understood within this context. After training/testing using Certified data, a model

may be applied to the remaining data for forming inferences. In other words, the model and its context may be extrapolated to infer MOOC non-certified use cases.

By comparing histograms from Certified data versus the entire dataset, we can see why segmentation is necessary. Figure 2 compares the histograms for grades of "ALL" (left) versus "Certified" (right). The y-axis scales are different since about 3% of courses taken are Certified. For "all data", the spike at zero defines the y scale.

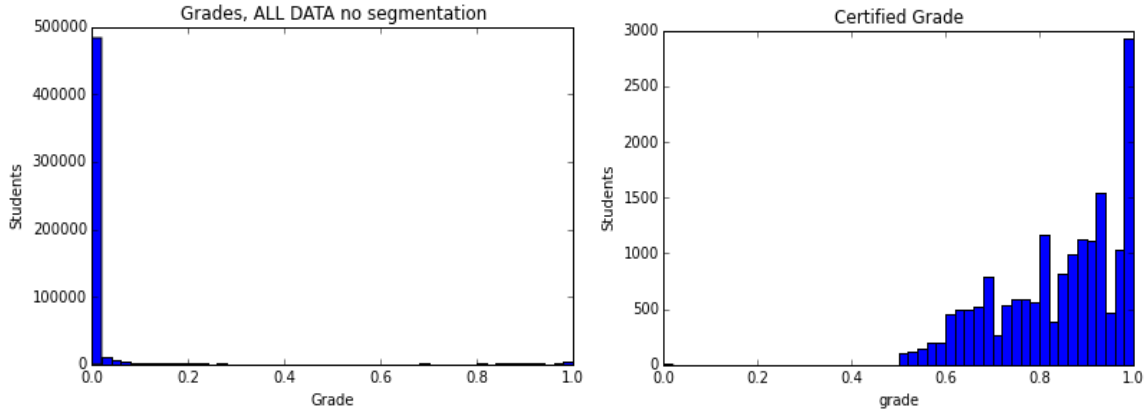


Figure 2: A histogram of grades for all courses taken (left) compared to a histogram of grades for Certified courses only (right).

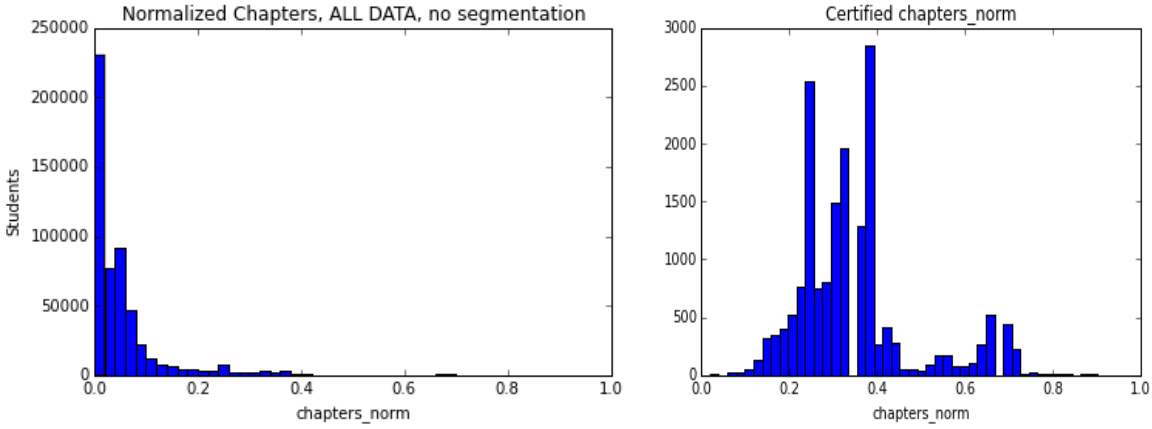


Figure 3: Normalized chapters histograms. All courses taken (left) and Certified only (right).

Comparisons of performance data are similar as shown in Figure 3 for normalized chapters. Again, the unsegmented dataset shows a very high population near zero while the Certified distribution is spread out with higher values. Figure 3 illustrates this difference for normalized chapters; the other performance histogram comparisons can be seen in Appendix A.

Figure 4 shows the histograms for "all data" and "Certified" courses only, for Level of Education of students taking courses. The shapes are nearly the same. Gender and Year of Birth histogram comparisons (see Appendix A) are similar, with no shifts as observed for participation data in Figure 2.

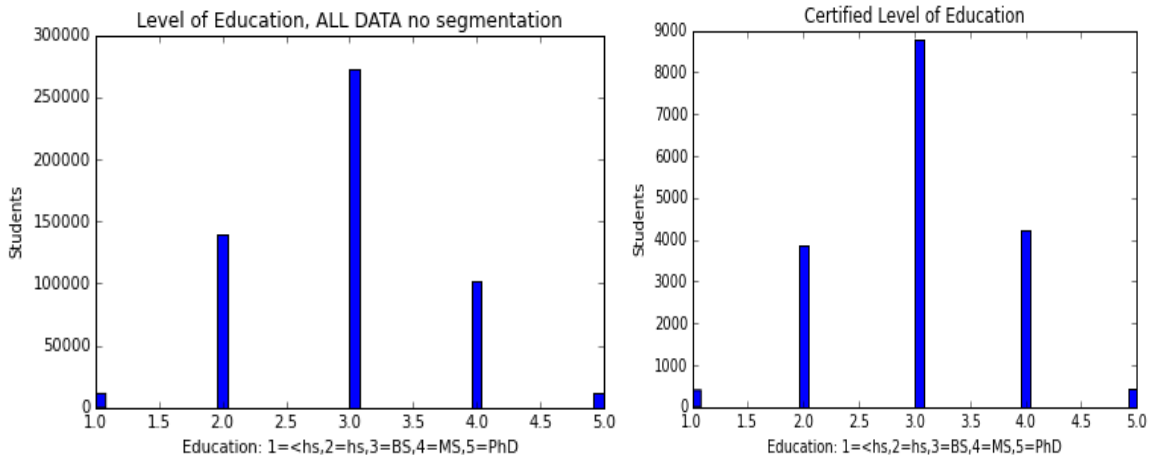


Figure 4: Level of Education histograms. All courses taken (left) and Certified only (right). The histogram shape is nearly unchanged between "all data" and "Certified".

2.3. EXTENDING THE TRAINING/TESTING DATA

At this point in processing, each row consists of numeric data only. There no longer is any student ID string or course title string, nor any other string type information.

Each row of X may be considered a vector x corresponding a y (grade) extending from that row. Within each row (vector) x, we can further consider that the variables are of two types: invariant student-based variables or student background information (gender, level of education and year of birth), and course participation variables (nevents, nvideo, nchapters_norm, etc). While the student-based data remains invariant, it is possible to further transform the participation (part of vector x) and grade (y) information for the purposes of training and testing a dataset designed for the non-Certified data.

The authors make an assumption based on the performance of the Certified segment, meaning students who successfully passed the course. Given similar participation characteristics for Auditor students to Certified students, the grade for the Auditor students would be comparable to those of the Certified students. For example, if an Auditor student's participation matched that of a Certified student halfway through the course, then the score of the Auditor student would be best represented by the Certified student's score at that halfway point. By modeling the Certified student's score based upon participation variables, a score can be calculated for all participation levels. This assumption allows the approximation of a grade for all Auditors at the time they drop out of the class.

In this manner, it is possible to construct a training dataset for a predictor, based on Certified course data, that spans the entire range from 0.0 to 1.0.

Certified data is "extended" to zero by creating X,y, where the participation vector x and y are multiplied by a random number between 0.0 and 1.0. This serves to "stretch" the dataset as noted above to cover the non-Certified domain. Figure 5 shows the original Certified grades (y) for

training (left) and the "extended" transform of these grades (right). It is worth pointing out that the test set of Certified data is similarly extended, so that training error and testing error may be measured (computed) over the entire grade range from 0.0 to 1.0. Figure 6 shows the same (original, right and extended, left) "y" grades for the testing set.

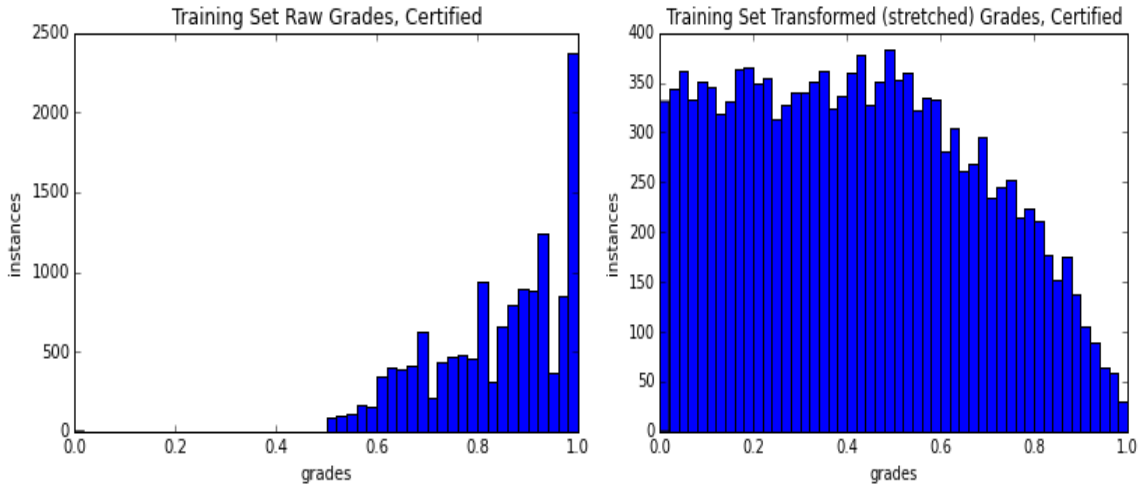


Figure 5: Training set grades (or "y") left, and extended training grades on the right.

The right side distribution is simply each datapoint multiplied by a uniform random number between 0 and 1, or the convolution of the left raw data with f(x') =1 for x' between 0 and 1, f(x')=0 otherwise. Histogram populations in the testing data are ¼ of those in the training data due to the 80%/20% split for machine learning.

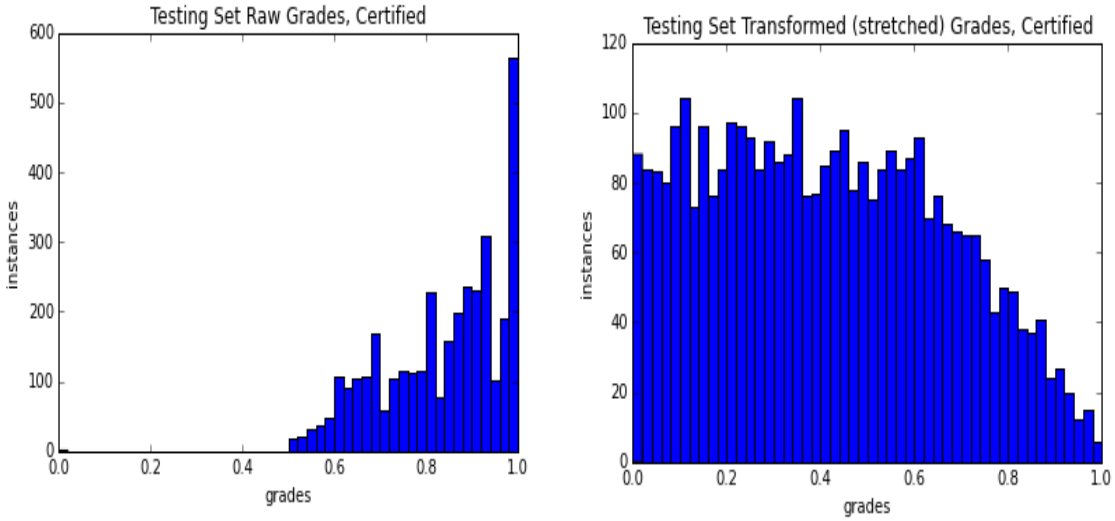


Figure 6: Testing set grades (or "y") left, and extended testing grades on the right.

2.4. MACHINE LEARNING MODELS UTILIZED

Predictive models utilized in this work include random forest (RF) regression and gradient boosted regression (GBR) packages.[18] GBR and RFR parameters were optimized to prevent overfitting[19] and minimize test error. RF and GBR models produced similar results. Parameters are shown in Figure 7 for an 80/20 split.

```
Random Forest:
from sklearn.ensemble import RandomForestRegressor as rf
reg = rf(n_jobs=-1, n_estimators=1000,max_features=1.0, min_samples_leaf=5,
            max_leaf_nodes=1000, max_depth=16, bootstrap=True)
Gradient Boosting:
from sklearn.ensemble import GradientBoostingRegressor as gbf
reg = gbf(n_estimators=400, max_features=0.75, min_samples_leaf=40,
max_depth=20, learning_rate=0.02)
```

Figure 7: Parameters for models which provided lowest testing errors.

2.5. TRAINING AND TESTING

Most of this machine learning work utilizes a split of 80% training data and 20% testing data. However, different splits (50/50, 60/40 and 90/10) provide nearly the same results as 80/20. This implies that the data size is sufficiently large since the smaller training sets did not show a higher level of training or testing error, and the largest size did not reduce these errors. Training and testing instances are chosen randomly with no "seed", so different runs use different combinations of training and testing instances. These different runs produce repeatable results, so cross validation has not been further explored.

Training and testing information is displayed in Figure 8. RF and GB testing rms error is about 0.098 and 0.097, respectively.

RANDOM FOREST REGRESSOR MODEL	GRADIENT BOOSTING REGRESSOR MODEL
training data accuracy: 0.927470910101	training data accuracy: 0.929317909476
test accuracy: 0.847456579928	test accuracy: 0.849576857812
training rms error 0.0683307209589	training rms error 0.0674550676204
testing rms error 0.0978013336966	testing rms error 0.0971192602659

Figure 8: Training and Testing Information for Regressor Models

Figure 9 shows correlation plots for RF (left) and GB (right) regressors. These plots show y_{pred}_test (predicted y, y axis) versus actual y_test (x axis). Predicted y values are generated by applying the trained models to X_test.

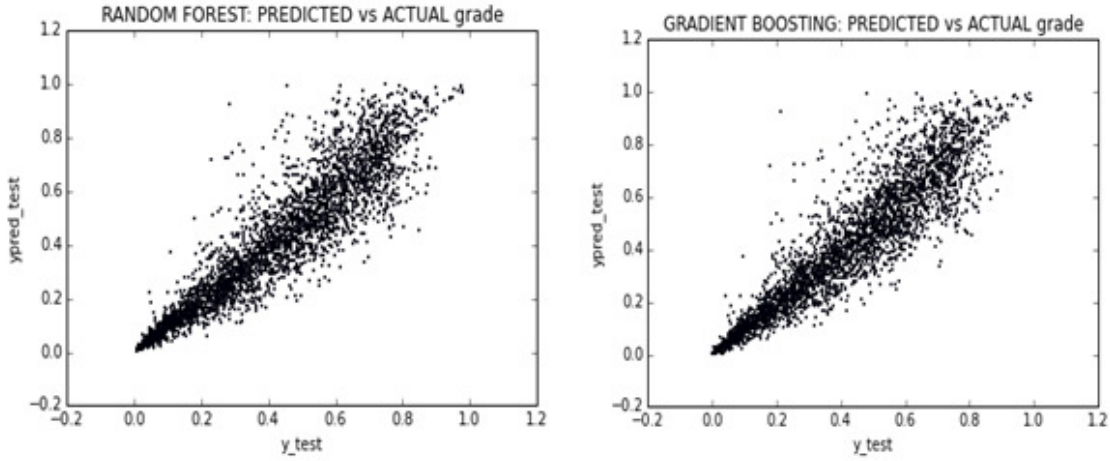


Figure 9: Correlation of y_predicted versus y for test data, for random forest regression (left) and gradient boosting regression (right). Point cloud shapes are about the same for both methods; resulting RMS error for both is about 0.098 and 0.097, respectively. An error of 0.1 (10%) is approximately a "traditional" A-B-C-D-F letter grade.

Additional validation work consisted of predicting the original test data and comparing histograms. The predicted result looks like a gaussian-convoluted (smoothed) version of the actual y. This was repeated for 'original' X,y test data multiplied 0.5, and separately by 0.25, in order to validate the models' extention of X,y data for to the grade range of 0.0 to 1.0.

2.6. Feature Importances

Figure 10 shows feature importances of the random forest model. Normalized chapters is the most important feature in the decision tree based models. Startend_days (calendar days between start and last event) and nevents (number of events) follow.

edX Feature Importances		Random Forest
Rank	Feature	Importance
1	chapters_norm	0.753
2	startend_days	0.143
3	nevents	0.0399
4	nplay_video	0.0224
5	ndays_act	0.0216
6	Year of Birth	0.0084
7	gender	0.0074
8	Level of Education	0.0032
9	nforum_posts	0.00097

Figure 10: Random forest regressor feature importances.

The least important features include student background (level of education, year of birth and gender). It is surprising that higher education does not show reduced participation for earning the same Certified grade. This contradicts previous inferernces[5].

2.7. Application To Non-Certified Data

From the predictive models applied to the non-certified student group, about 20,000 students are in the category which we may call Auditors. These students have viewed the entire course,

indistinguishably from the Certified group, but have not produced graded work. This is similar to a student who attends every lecture in a course, who sits in the front row taking notes, so is indistinguishable from other students' course behavior except they do not take tests or hand in homework for a grade.

In order to predict auditor status, the non-certified data (X) is fed into the random forest (and gradient boosting) regressors, with similar results shown in Figure . For this processing, the non-certified ground-truth "y" is ignored. The model has been trained using the extended Certified data, and it is understood that the actual grades are mostly near zero, always below the cutoffs for Certified status.

This analysis assumes that the online accessing of available resources leads to learning and may not be applicable to graduate level courses for which difficult assignments and independent research comprise the greatest part of a student's investment in time and struggle in the acquisition of skills and learning of key concepts. Nevertheless, we may measure participation in terms of the students who were certified to better understand the use cases for MOOCs. Since the overwhelming majority of courses taken do not reach the Certified status, the majority of resource use pursues non-certified learning.

To simplify plots, students with all-zero participation are removed from the data. Approximately 100,000 instances fall into this zero-participation category. These map into traditional students' unregistering prior to starting a semester.

Figure 11 describes the distribution of predicted grades for random forest and gradient boosted regressors. Compared to actual grades in Figure 2 (left), predicted grades are shifted to larger values, showing that non-certified users are likely to skip graded participation. Auditors are defined as those students with a predicted grade greater than or equal to the minimum grade required for Certified status (typically 0.6). Figure 12 shows the auditor histograms (effectively Figure 9 zoomed in to the 0.6 to 1.0 range).

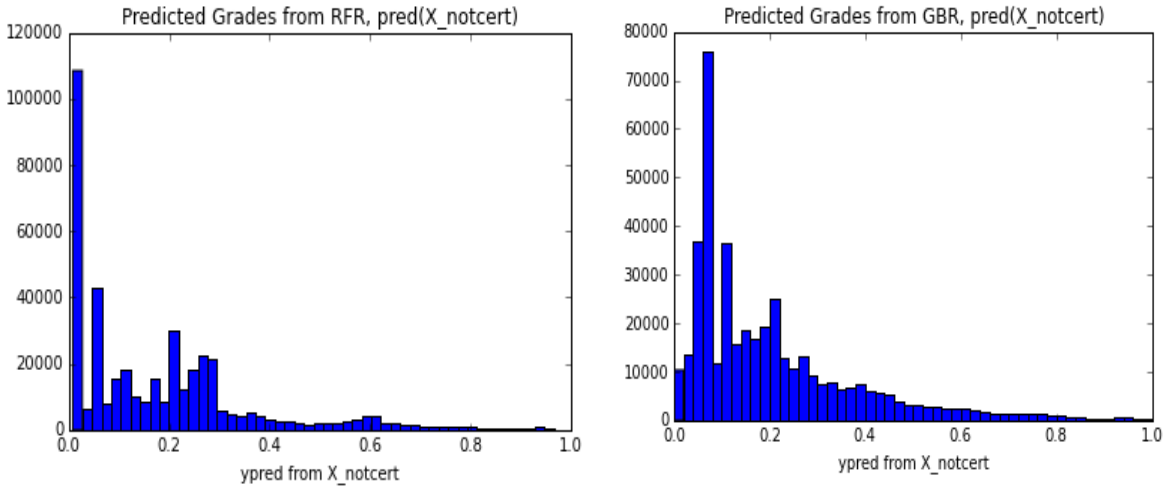


Figure 11: Predicted grades from Random Forest regression (left) and Gradient Boosted regression (right). The histograms represent predicted grades for 420,000 courses taken (without the certified ~18K, zero ~100K removed from the original, clean ~540K instances).

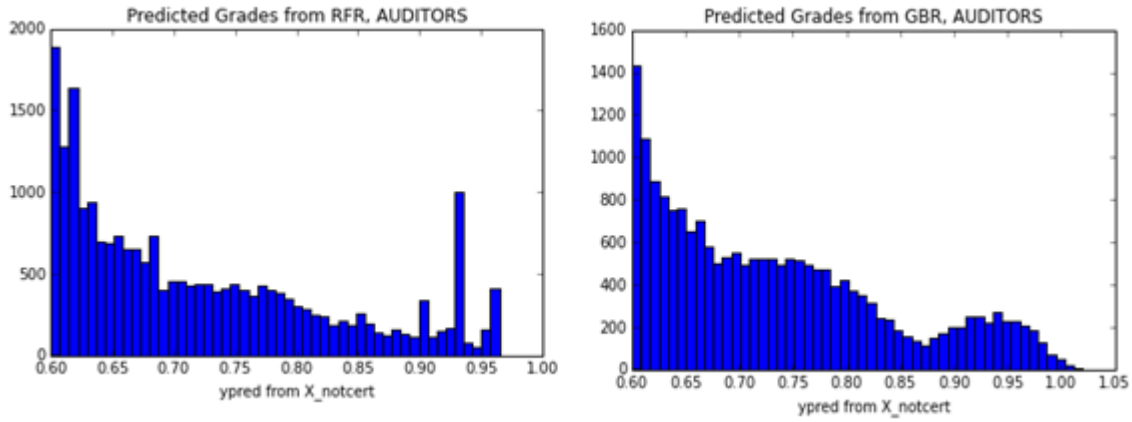


Figure 12: Predicted grades for Auditors, RFR (left) and GBR (right). Random forest: 23256 auditors; gradient boosting: 20078 auditors. Histograms and number of auditors are similar.

The auditor group (20,000 to 23,000) is larger than the Certified group (~18,000). The distribution of auditors shows the greatest density on the low end, at the hard-boundary for auditing (>=0.6 predicted grade). The displayed auditor histograms are decreasing tails of distributions whose peaks are closer to zero.

What is the mean non-certified predicted grade? What is the cumulative effort for the non-certified courses, in terms of Certified efforts?

Mean Certified Grade: 0.8356
Mean Non-Cert Grade: 0.0119
Mean PREDICTED Non-Cert Grade: 0.1996
Non-Cert Courses Taken (excludes zeros): 421,000
Non-Cert Participation in terms of Cert Mean Course: 100,500

Figure 13: Certified and Non-Certified course Stats

Figure 13 shows statistics related to non-certified predicted grade and participation. If we measure participation with predicted grade, the cumulative participation for non-certified courses is the equivalent of 100,000 certified courses (about 6X the actual certified courses achieved). The efforts of students who do not complete courses form a substantial part of total efforts expended in using course materials, and their use cases should not be ignored by course organizers.

Summary: Auditors form a group which is larger than the Certified courses completed with assignments earning high grades. Most use of course material is made in non-certified study.

3. General Application Of Analysis

The analysis described above may be applied generally to other (non-educational) analyses. For example, it may be possible to collect data for online purchasers of a given product. The actions of the purchasers are analogous to the performance metrics of Certified students. It may be possible to determine use cases such as “window shoppers” or “aficionados” by comparing the (small but significant) purchasing customer segment to the remaining majority of non-purchasers.

The analysis described above may also be applicable to other behavior-related, segmentable populations to study phenomena such as auto accident occurrence.

4. CONCLUSIONS

Regression models explain grades based on participation variables, thereby reducing unexplained grade variance by 70%. Random forest and gradient boosting provides the least error in testing of regression models.

We have found the "Auditor" use case within non-Certified data. In this paper, Auditors have been defined as those students whose course participation metrics are indistinguishable from those who complete a course with a Certified status and high grade. The auditing group (20,000+) is larger than the Certified group (17,700). Auditor use of online resources is approximately at the same level as Certified use.

Student background is irrelevant to prediction of grades. These variables are among those with the lowest feature importance; removing student participation from the dataframe has no effect on model error. Furthermore, a model using only student background (without participation) results in unexplained variance equal to that of the naive model.

REFERENCES

[1] Deconstructing Disengagement: Analyzing Learner Sub-populations in MOOCs, Kizilcec, R.F., Piech, C., Schneider, E. 2013 http://web.stanford.edu/~cpiech/bio/papers/deconstructingDisengagement.pdf

[2] HarvardX and MITx: Four Years of Open Online Courses -- Fall 2012-Summer 2016, Isaac Chuang, Andrew Dean Ho, December 2016,https://papers.ssrn.com/sol3/papers2.cfm?abstract_id=2889436

[3] Diversity in MOOC Students' Backgrounds and Behaviors in Relationship to Performance in 6.002x, Jennifer DeBoer, Glenda S. Stump, Daniel Seaton, Daniel Seaton, 2013. http://tll.mit.edu/sites/default/files/library/LINC%20'13.pdf

[4] HarvardX and MITx: The first year of Open Online Courses, HarvardX and MITx Working Paper #1, https://journalistsresource.org/wp-content/uploads/2014/01/SSRN-id2381263.pdf

[5] Who Performs Best in Online Classes?, Max Woolf, @minimaxir, July 10, 2014.http://minimaxir.com/2014/07/online-class-charts/ edX

[6] Bioelectricity: A Quantitative Approach Duke University's First MOOC Feb 5, 2013, http://dukespace.lib.duke.edu/dspace/bitstream/handle/10161/6216/Duke_Bioelectricity_MOOC_Fall 2012.pdf

[7] http://mfeldstein.com/emerging_student_patterns_in_moocs_graphical_view/Phill Hill, e-Literate, March 6, 2013 (describes MOOC user types)

[8] https://hbr.org/2015/09/whos-benefiting-from-moocs-and-why Harvard Business Review

[9] HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0, (dataset .csv file, .pdf file explaining variables, and .pdf file explaining de-identification) https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147

[10] https://www.insidehighered.com/news/2016/11/02/georgia-institute-technology-award-credit-through-massive-open-online-course Undergrad online classes thru MOOC's

[11] http://www.skilledup.com/articles/the-best-mooc-provider-a-review-of-coursera-udacity-and-edx

[12] http://moocnewsandreviews.com/who-should-take-a-mooc-9-types-of-lifelong-learners-who-can-benefit/

[13] http://project.ecolearning.eu/mooc-types/

[14] https://journalistsresource.org/studies/society/education/moocs-online-learning-research-roundup

[16] New model for measuring MOOCs completion rates, Syed Munib HADI & Phillip GAGEN, EMOOCs 2016, European Stakeholders Summit http://www.academia.edu/20434486/New_model_for_measuring_MOOCs_completion_rates

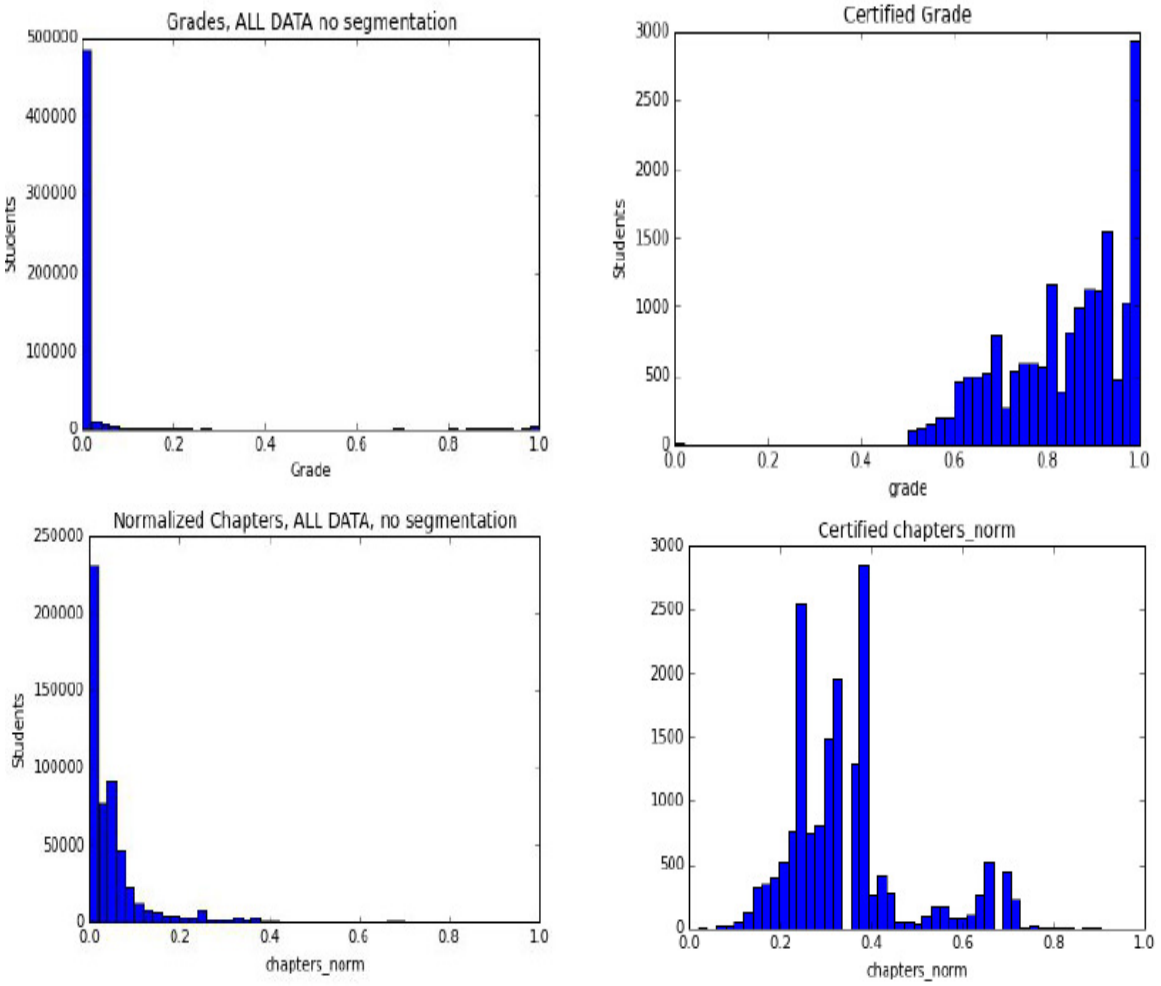
[17] Greene, Jeffrey A., Christopher A. Oswald, and Jeffrey Pomerantz. "Predictors of retention and achievement in a massive open online course." American Educational Research Journal (2015): 0002831215584621.http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.908.4811&rep=rep1&type=pdf

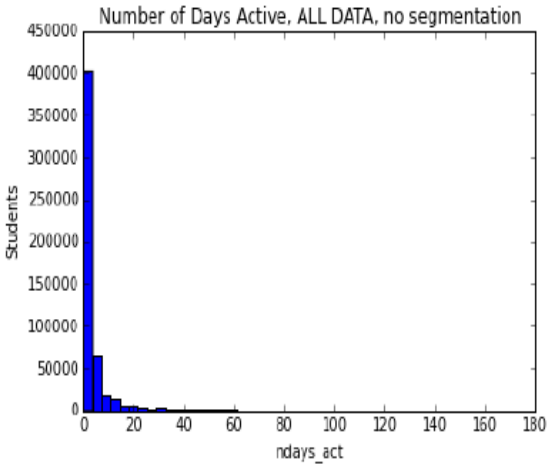
[18] Sckikit-learn documentation http://scikit-learn.org/stable/user_guide.html

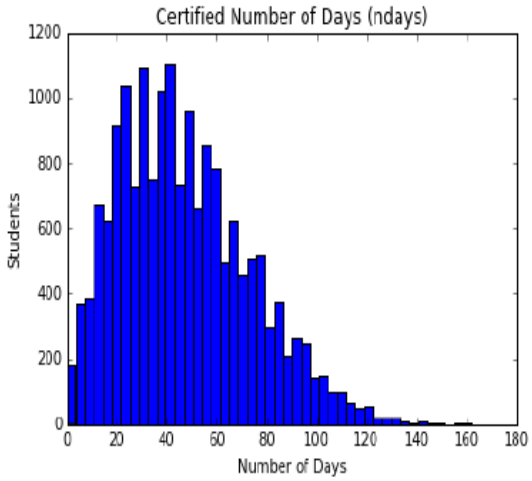
[19] Machine Learning, p. 70, Tom M. Mitchell, 1997, Mcgraw-Hill.

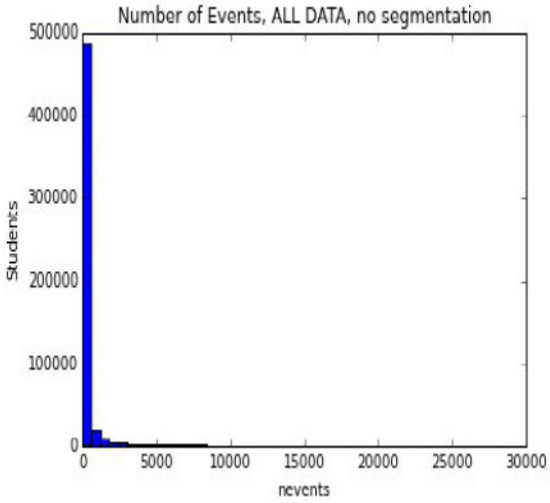
APPENDIX A:

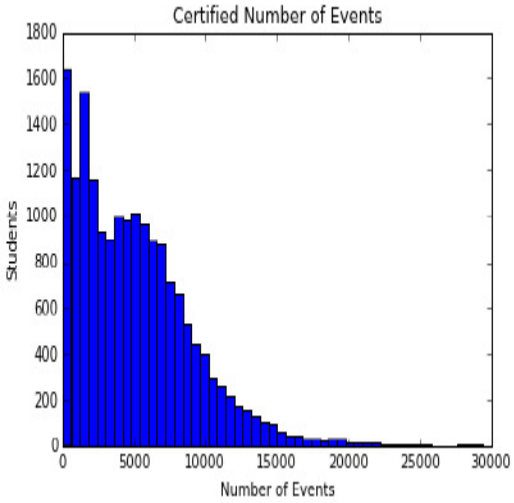
Histograms comparing All Course Data versus Certified Course Data

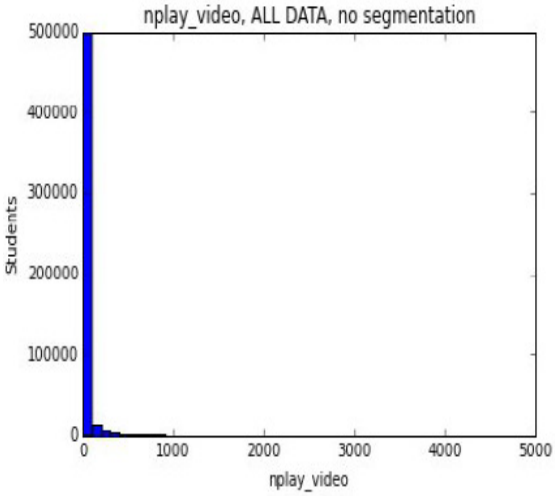


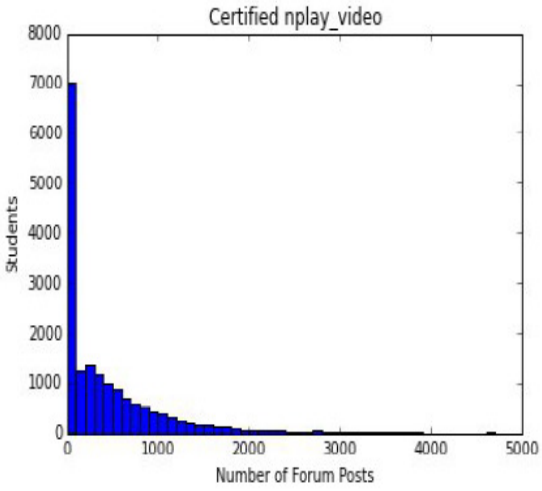

Number of Days Active, ALL DATA, no segmentation
Students
ndays_act

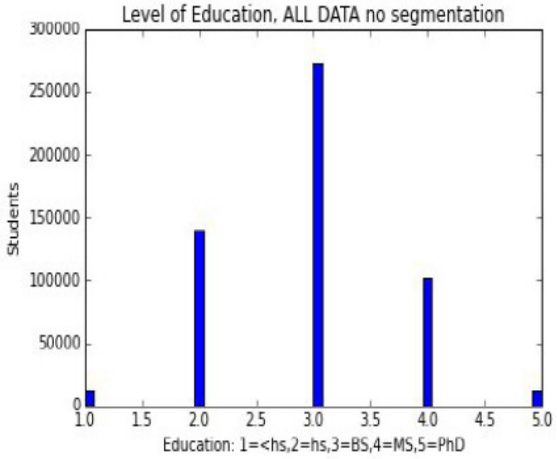

Certified Number of Days (ndays)
Students
Number of Days

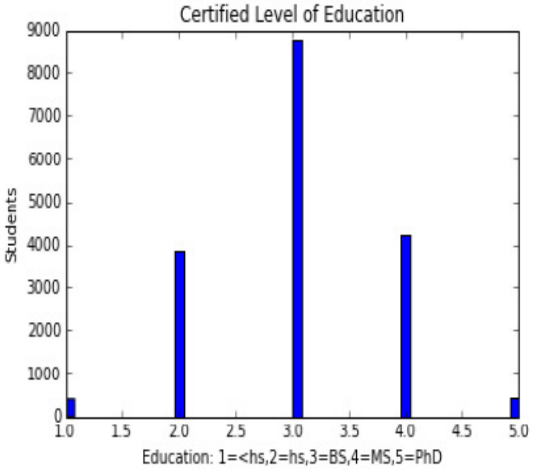

Number of Events, ALL DATA, no segmentation
Students
nevents

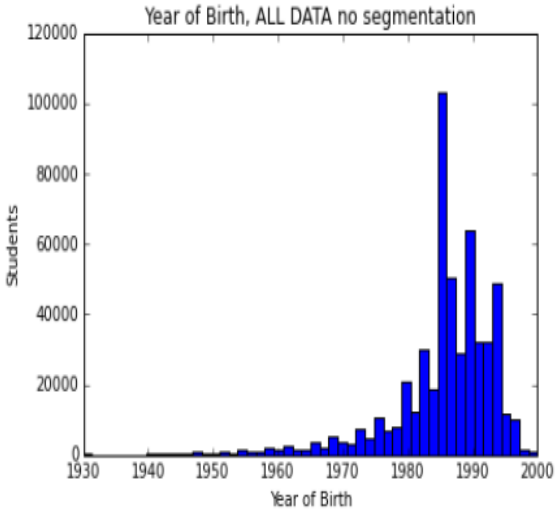

Certified Number of Events
Students
Number of Events

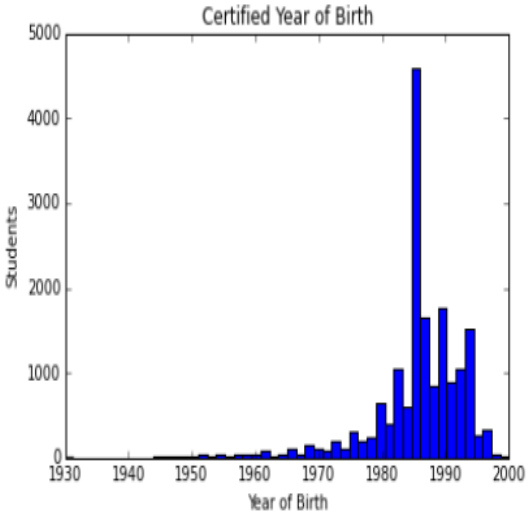

nplay_video, ALL DATA, no segmentation
Students
nplay_video


Certified nplay_video
Students
Number of Forum Posts


Level of Education, ALL DATA no segmentation
Students
Education: 1=<hs,2=hs,3=BS,4=MS,5=PhD


Certified Level of Education
Students
Education: 1=<hs,2=hs,3=BS,4=MS,5=PhD


Year of Birth, ALL DATA no segmentation
Students
Year of Birth


Certified Year of Birth
Students
Year of Birth