Title
A Pilot Study of Predicting Failing Grades Using Data from UCLA's Learning Management System

Author
Kang, Elliot

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

A Pilot Study of Predicting Failing Grades Using Data from UCLA's Learning Management System

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Elliot Minkiu Kang

2017

ABSTRACT OF THE THESIS

A Pilot Study of Predicting Failing Grades
Using Data from UCLA's Learning Management System

by

Elliot Minkiu Kang
Master of Science in Statistics
University of California, Los Angeles, 2017
Professor Robert L. Gould, Co-Chair
Professor Mark Stephen Handcock, Co-Chair

UCLA develops and uses a learning management system to provide an online environment for students to access and interact with course content. The data collected by the learning management system is a direct measure of student activity, and provides information that augments known information about the student, such as assignment grades and demographics. This paper assesses UCLA's learning management system data for its usefulness in creating an early warning system that will advise instructors and students of whether a student is likely to receive a failing grade. The data are used in two analyses: an exploratory analysis of how students use the learning management system, and a predictive model to forecast end-of-term grades based on partial-term information. Recommendations on how to generalize the results across the UCLA undergraduate population are drawn from the findings of the analyses.

The thesis of Elliot Minkiu Kang is approved.

Kathleen L. Komar

Chad J. Hazlett

Mark Stephen Handcock, Committee Co-Chair

Robert L. Gould, Committee Co-Chair

University of California, Los Angeles

2017

Ps. 115:1.

Rom. 11:36.

Rev. 5:9-14.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

As I was thinking about what I wanted to research for my thesis, Laura Tao told me about an idea that had been proposed to her. She was interested in other topics, and passed the idea on to me. That idea turned into my topic for this thesis. I'd like to thank Laura not only for introducing the idea of studying CCLE data, but also for the moments where we have been able to catch up and commiserate over some of the difficulties of life in graduate school.

David Baron and Janet Yee from Southern California Gas Company were kindly willing to let me use some of their data for my thesis. Although I ended up choosing different data to study, I appreciate their investment in me over the last couple of years, culminating in their offer of data.

Michelle Lew and Nick Thompson from the Office of Instructional Development, and Mark Levis-Fitzgerald and Hannah Whang Sayson from the Center for Educational Assessment graciously took their time to meet with me and plan how I could obtain the data I needed for my thesis. Michelle and Mark provided some early guidance for variables they thought would be interesting, and informed me of administrative hurdles I would need to clear. Nick and Hannah performed the data extracts and anonymization. All of this was helpful in navigating the unknown (to me, at least) territory of UCLA data procedures.

Professor Rob Gould took time to meet with me regularly to advise my thesis. I am grateful for his help during the entire research process, from how to get the data, to suggestions of analyses to try. His reassurances and encouragement kept me going through the longest analysis I've ever completed. Glenda Jones answered all of my questions–and there were a lot of them. She patiently helped me form my committee, add a member to it, then advance to candidacy. (Yes, in that order.)

I recognize more every year the love displayed by a parent, or a sibling. Thank you, Mom, Dad, and Ryan for your encouragement and support.

Grace on Campus reminds me every week of what is most important: "Fear God and keep his commandments, for this is the whole duty of man" (Ecc. 12:13b). And old 301–Ryan, Seichi, Andrew, and Jon: I am still affected by echoes of conversations we had when we lived together. I am thankful for our friendship.

Finally, to God: "For Christ also suffered once for sins, the righteous for the unrighteous, that he might bring us to God" (1 Pet. 3:18a)–there is no truth more precious. "For from him and through him and to him are all things. To him be glory forever. Amen" (Rom. 11:36)!

CHAPTER 1

Introduction

UCLA has adopted and developed an online learning management system (LMS), as a member of the collaborative group that maintains Moodle, to facilitate student learning. This system, called "Common Collaboration & Learning Environment" or "CCLE" [6], provides course instructors and teaching assistants with websites with a wide array of features. These features include uploading syllabi and other course-related files, posting links, sending announcements, hosting forum discussions, receiving homework submissions, and assigning quizzes [13].

This thesis focuses on using LMS-collected metadata about student online activity to assess whether there are any discernible differences in how students at UCLA use CCLE. Moreover, a preliminary prediction algorithm will be created that assesses how likely it is, based on information from the first half of the term, that a student will fail the course. This thesis acts as a pilot study to determine whether UCLA could utilize a predictive model to alert students and course instructors via CCLE of whether a student will fail.

Moodle, a collective open source development platform that develops the LMS for CCLE, advertises that it "delivers a powerful set of learner-centric tools and collaborative learning environments that empower both teaching and learning" [12]. Other LMS vendors tout that "[p]ersonalized learning solutions... enhance personalized and competency-based education and increase student engagement" [3] and that "[e]very last feature, every last interface is crafted... to make teaching and learning easier" [5]. One of the primary selling points for LMSs is that they facilitate and improve learning.

Indeed, LMSs can be helpful in a college environment where 87% of college students use a laptop every week to do schoolwork, and 85% of college students own a smartphone [18].

As of 2011, an estimated 88% of undergraduate and 93% of graduate students owned laptops [7]. There are great potential benefits to utilizing a LMS in a college setting, such as allowing students to download lecture content such as handouts and presentation slides, so that they can review the source material outside of the classroom.

In addition to the beneficial services that LMSs provide, the data they generate are also another valuable resource. Every interaction that a student has with the LMS, such as downloading files and clicking links, can be stored and analyzed; LMSs are also capable of storing assignment grades. These two sources of data together–student activity and grades–make a LMS a robust repository of student data. Notably, student interaction with a LMS is not immediately observable to the instructor, and thus these data provide measurements of student activity that are entirely complementary to an instructor's other measurements of student engagement. These LMS data can be aggregated for simple reports, such as the number of student submissions for an assignment, or the average score on an online quiz. But the data can also be leveraged for more complex analyses, such as clustering and predictive modeling.

One particular use of LMS data is for grade prediction, for which there is a clear and immediate benefit: if the course instructor is able to anticipate in advance of submitting the final grade–perhaps halfway through the term–that a student is likely to fail, he or she will be able to intervene early in providing these students with additional resources to help them learn and master the course content. While some useful data are already available to the instructor, such as grades on assignments like homework and exams, LMS data provides direct measures of student engagement that augment measures of student performance, like graded assignments.

This thesis is a pilot study that utilizes one course from CCLE for analysis. Data have been collected about the STATS 10 course taught by Professor Rob Gould in the winter term of 2016, to see if it is feasible to create a model to predict whether a student will fail the course based on information available during the middle of the term. If the model is satisfactory, the model can be tested using data from other undergraduate courses, and other models can be built to handle differences among those courses.

In this thesis, Chapter 2 describes the three sources of data used: UCLA CCLE data provided by the Office of Instructional Development, UCLA Registrar's Office demographic data provided by the Center for Educational Assessment, and graded assignment data provided by Professor Rob Gould of the Department of Statistics. Chapter 3 presents descriptive statistics and analyses for identifying patterns of how students at UCLA use CCLE. Chapter 4 provides preliminary models that can be used to predict final grades, and compares them across different options for input data sources and outcome variables, as well as against a baseline predictive model. Chapter 5 offers thoughts on generalizing the findings, and Chapter 6 concludes.

CHAPTER 2

Overview of Data

The data for this thesis are owned by various offices at UCLA, and were obtained after a request for access was submitted to the UCLA Institutional Review Board (IRB). Protocols were proposed and approved by the UCLA IRB to ensure the confidentiality of the students who generated the data. Although university-assigned ID numbers are unique identifiers and thus allow the joining of multiple sources of data, their inclusion creates a concern for student confidentiality. Thus before all the data were received, the university-assigned ID numbers were replaced with anonymized unique identifiers. This still allowed for data joining, but preserved confidentiality. Another step to maintain confidentiality was to exclude from transfer any demographic variables where one subgroup was 10% of the sample or smaller. This was to mitigate the potential to identify individuals by uncommon demographics. (Exceptions were allowed if the group of 10% or smaller was labeled as "unknown.")

For this thesis, the data were from a STATS 10 course taught by Professor Rob Gould in the winter term of 2016, starting on January 4 and lasting for 11 weeks. 150 students completed the course and received a final grade. Because this analysis is a pilot study, the scope of the data is small; were CCLE data to be used in a predictive model for the undergraduate population, a much wider scope of data would need to be used so that the data used would resemble the courses served by the predictive model.

The CCLE LMS data come from the UCLA Office of Instructional Development (OID). The data are stored in a relational database, and several individual tables were exported as flat files. Of the tables provided, the one used for analysis was a log of every action in the CCLE course website. A sample observation from the log can be seen in Table A.1.

In this log file, each row stores information from every time a user clicked a link within CCLE, such as to open a page or download a file. (Hereafter, "interaction" or "event" will also be used to describe one row of the log.) In order to create a wide dataset for analysis, where each row represents a user, rather than an interaction, the data were cleaned: all administrative actions by the CCLE system were removed, and actions from outside the term were excluded. (The term's start and end dates were retrieved from the UCLA Academic Calendar [17].)

In the cleaning process, new variables were created from the activity data: total number of interactions, number of sessions (where a new session is defined as occurring at least one hour after the last event), average session length, average time between sessions, the number of times the syllabus was downloaded, the number of times a file was downloaded, the number of times the announcement forum was opened, the number of times an announcement post was read, the number of times the discussion forum was opened, and the number of posts to the discussion forum. (The announcement forum is for posts by instructors and instructional assistants only, whereas the discussion forum is for posts by anyone able to access the course page.) These variables were chosen for their ease of creation from the data.

Demographic information was provided as a capture of Registrar's Office data by the Center for Educational Assessment (CEA). The data contain student demographics at the end of the 2016 winter term, not at the time of analysis. Thus despite student demographics that change, such as year in school and grade point average, the demographics data regard the term being studied. This prevents more current data from being anachronistically incorporated into the analysis. Not all demographic variables were used; those included in analysis are the first registered term, transfer status, sex, ethnicity, residence status, first generation student status, and Pell Grant recipient status. These variables were chosen for being common demographic markers of students.

Student assignment grades were provided by Professor Rob Gould in an Excel file, which was pre-processed in Excel so that each column would represent an assignment grade. The data were then cleaned so that missing grades were given a score of 0, and excused grades

were given the student's final grade. A full list of assignments and their due dates can be seen in Table 2.1.

Assignment	Due Date	Week Number
Homework 1	1/8/2016	1
Homework 2	1/15/2016	2
Lab 1	1/19/2016	3
Homework 3	1/20/2016	3
Exam 1	1/25/2016	4
Homework Quiz 1	1/26/2016	4
Lab 2	1/26/2016	4
Homework 4	1/29/2016	4
Homework 5	2/5/2016	5
Lab 3	2/12/2016	6
Homework 6	2/16/2016	7
Homework 7	2/19/2016	7
Exam 2	2/19/2016	7
Lab 4	2/25/2016	8
Homework 8	2/26/2016	8
Homework 9	3/4/2016	9
Homework Quiz 2	3/8/2016	10
Lab 5	3/10/2016	10
Homework 10	3/11/2016	10
Final Exam	3/16/2016	Finals
Participation	3/21/2016	Finals

Table 2.1: All assignments with dates and weeks; the line indicates assignments due in the first five weeks.

2.1 Slicing CCLE and Assignment Data by Week

Both the CCLE and assignment data contained temporal information. CCLE logged the timestamp for each event, so each record had a date-time entry. The assignment data came with due dates, so each assignment grade is linked to a date. The inclusion of date information allows the analysis to use data only from before a chosen date; this simulates an analysis being run in the middle of the term, only knowing part of the data.

A caveat is that the dates in the assignment data are the due dates, and not the dates when the grades were entered; the dates when the grades were entered are unknown. This means that grades for an assignment may only have been known the week after the assignment was due. For example, an assignment with a due date of February 12, 2016 (the Friday of week 6) may have had grades entered on February 19, 2016 (the Friday of week 7). In such cases, the grades for that assignment should not be included in a model using information only up to the week the assignment was due, since the grades were unknown at that time. But because the only dates available are the due dates, these will be used as a proxy for when the assignment grades were entered, and subsequently, when that information became available as input data for a model.

2.2 Formatting Grades as the Outcome Variable

Students' final letter grades were obtained through the demographics data from CEA. As provided, the grades were provided in half-letter format (e.g. A+, A, A-, etc.), as well as with values for "passed" (P) and "not passed" (NP).

To create different options for the predictive model, two other outcome variables were created from the half-letter final grades. First, full-letter grades were created by dropping the "+" and "-" suffixes from each letter grade. Because a "P" grade indicates "achievement at grade C level or better" [16], all values of "P" were changed to a "B" (the average passing full-letter grade) and all values of "NP" were changed to a "D" (the average non-passing full-

letter grade). Then pass/fail grades were created with A-level, B-level, and C-level grades collapsed to "passing," and the D-level and F-level grades collapsed to "failing."

The following table shows how half-letter grades were converted to full-letter and passed/not passed (or "pass/fail") grades:

Half-letter	A+	A	A-	B+	B	B-	P	C+	C	C-	D+	D	D-	NP	F
Full-letter	A			B				C			D				F
Pass/fail	Pass										Fail				

Table 2.2: Scheme for half-letter grades to full-letter and pass/fail grades.

CHAPTER 3

Exploratory Analysis

3.1 Descriptive Statistics

There are 150 students who received final grades in the course. Distributions of half-letter, full-letter, and pass/fail grades can be seen in Tables 3.1 to 3.3.

A+	A	A-	B+	B	B-	C+	C	C-	D+	D	F	P	NP
4	38	8	15	23	8	12	25	2	1	2	3	5	4

Table 3.1: Distribution of half-letter grades, with pass and no-pass values.

A	B	C	D	F
50	51	39	7	3

Table 3.2: Distribution of full-letter grades.

Passed	Failed
140	10

Table 3.3: Distribution of pass/fail grades.

The demographics for the course are similar to those for the entire UCLA undergraduate population. By inspection, the class proportions for sex, residence status, and ethnicity appear not to be too different from the proportions for UCLA.

	Count	Proportion	UCLA
Male	61	40.67%	43.3%
Female	89	59.33%	56.7%

Table 3.4: Students by sex.

	Count	Proportion	UCLA
Freshman	132	88%	76.3%
Transfer	18	12%	23.7%

Table 3.5: Students by transfer status.

	Count	Proportion	UCLA
Resident	117	78%	73%
Non-resident	33	22%	27%

Table 3.6: Students by residence status.

	Count	Proportion	UCLA
Asian or Pacific Islander	41	27.3%	32.1%
Hispanic or Black	49	32.7%	25.7%
White Non-Hispanic	36	24%	26.3%
Foreign or international	15	10%	11.9%
Unstated, Unknown, Other	9	6%	3.5%

Table 3.7: Students by ethnicity.

	Count	Proportion
First-generation	64	42.7%
Not first-generation	81	54%
Unknown or missing	5	3.3%

Table 3.8: Students by whether a parent completed a bachelor's degree.

	Count	Proportion
Non-recipient	87	58%
Recipient	63	42%

Table 3.9: Students by Pell Grant status.

UCLA publishes student population demographics for sex [19], transfer status [19], residence status [20], and ethnicity [1]. χ^2 tests of independence for sex, residence status, and ethnicity show independence between the variable and whether the student population of interest is STATS 10 or UCLA (p-values of 0.730, 0.383, 0.523 respectively); however, transfer status is dependent (p-value of 0.013). (To obtain counts for UCLA, the proportions were multiplied by the number of students enrolled in STATS 10. The χ^2 test of independence was then performed on the counts for both groups.)

A likely reason for STATS 10's resemblance of the UCLA undergraduate population is that it fulfills the general education (GE) requirement for most undergraduate students [15], thus encouraging a broad cross-section of students across the undergraduate population to enroll in the course. The notable difference between STATS 10 and the UCLA undergraduate population is transfer status: STATS 10 had a higher proportion of freshman admits than UCLA did. Because STATS 10 is a GE course, transfer students may be likely to take an equivalent course earlier in their academic career, at their earlier institution.

Two-way tables begin to indicate associations between variables in the data. In particular, in view of the goal to predict student grades, two-way tables involving student grades suggest some association between grades and demographic variables:

	A	B	C	D	F
Asian or Pacific Islander	22	11	6	1	1
Hispanic or Black	4	20	21	3	1
White Non-Hispanic	17	12	5	1	1
Foreign or international	5	8	1	1	0
Unstated, Unknown, Other	12	0	6	1	0

Table 3.10: Students' grades by ethnicity.

	A	B	C	D	F
Non-recipient	36	31	15	3	2
Recipient	14	20	24	4	1

Table 3.11: Students' grades by Pell Grant recipient status.

χ^2 tests of independence on Tables 3.10 and 3.11 yield p-values of 0.0003 and 0.026 respectively. Because some of the expected counts are fewer than 5, the tests only approximate the p-values. Using a pass/fail grade scheme instead of full-letter grades does not alleviate the small expected counts, nor does removing the smaller ethnicity subgroups ("Foreign or international" and "Unstated, Unknown, Other"). Visual inspection of the tables plotted as histograms illustrates some association between the variables:

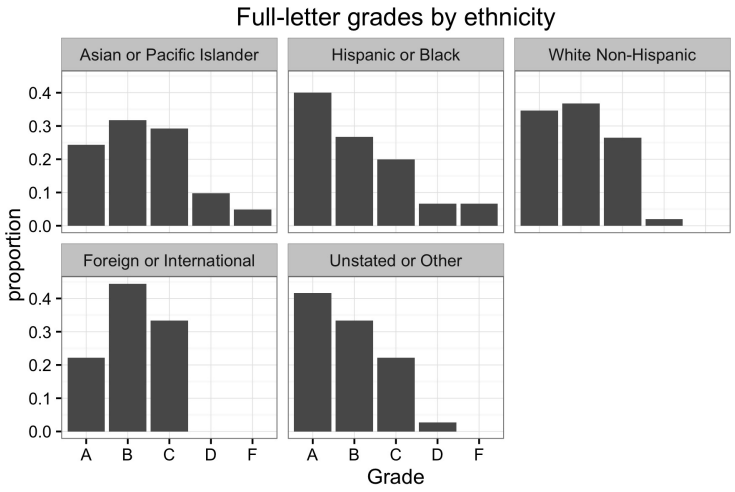

Figure 3.1: Full-letter grades by student ethnicity (each point represents a student).

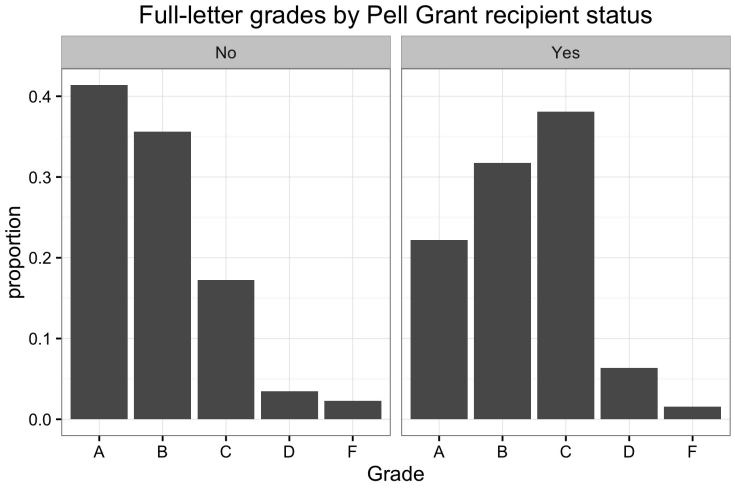

Figure 3.2: Full-letter grades by Pell grant recipient status (each point represents a student).

The CCLE data give an idea of how students tend to use the website. Figures 3.3 to 3.5 plot two general measures of online activity via the number of clicks. (Other variables can measure the purpose of a click, such as whether a file is downloaded, or a forum is opened. The number of events and the number of sessions are blind to a student's intention behind an interaction, and simply measure when a student interacts with CCLE.) Both graphs are unimodal with a moderate right skew, indicating that there are some students who have a high level of CCLE activity. Although the correlation between the number of events and

the number of sessions is high ($r = 0.82$), this simply means that the more sessions someone has, the more events they tend to have. This is likely due to the way that the number of sessions was constructed–as a function of the number of events, defining a new session as an event that occurs at least one hour after the last event. Thus it is expected that more clicks leads to more sessions.

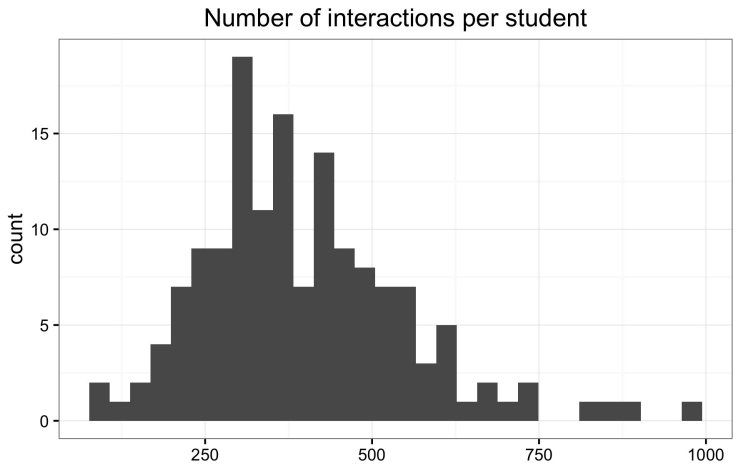

Figure 3.3: The number of interactions with the course CCLE page by student (each point represents a student).

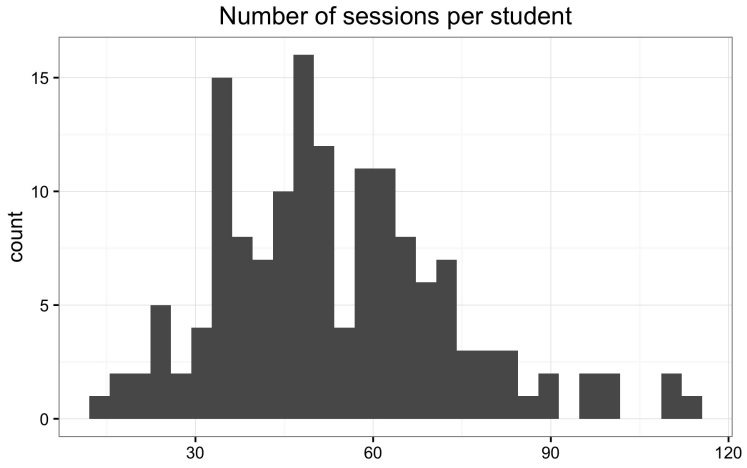

Figure 3.4: The number of sessions on the course CCLE page by student (each point represents a student).

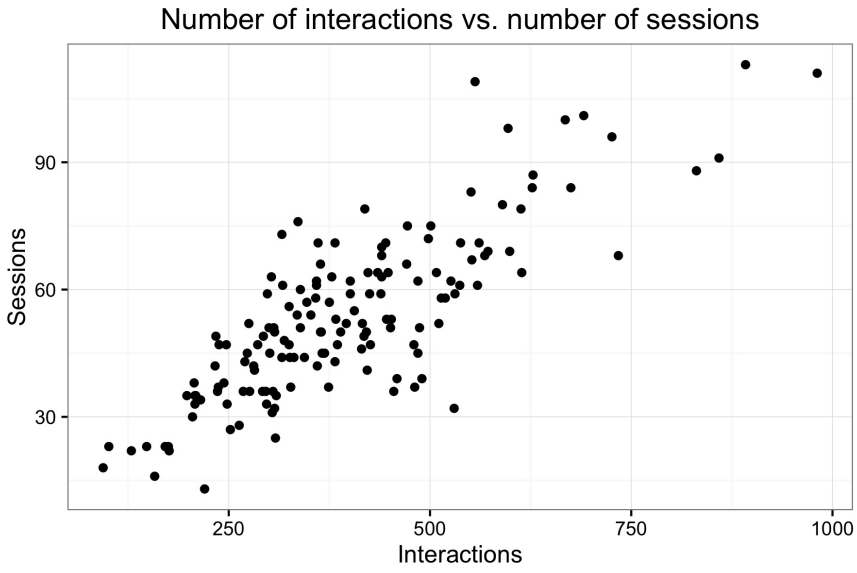

Figure 3.5: The number of interactions with the course CCLE page versus the number of sessions.

But the strong positive correlation between the number of interactions and the number of sessions could occur in two ways: with high activity density or with low activity density. In the former, each session would contain a high number of clicks, and so the two would increase together quickly, as the density of clicks within each session would be high. In the latter, each session would contain a moderate to lesser number of clicks, so although the two would increase together, the density of clicks within each session would be lower. There is a weak negative correlation ($r = -0.24, p = 0.003$) between the average number of events per session and the number of sessions, indicating that a greater number of sessions is actually associated with a slightly lesser activity density. For this course, students who use CCLE more, as measured by the number of events, tend to have fewer clicks in a session.

Examining sessions in terms of time instead of events, both the average session length and the average time between sessions are right-skewed:

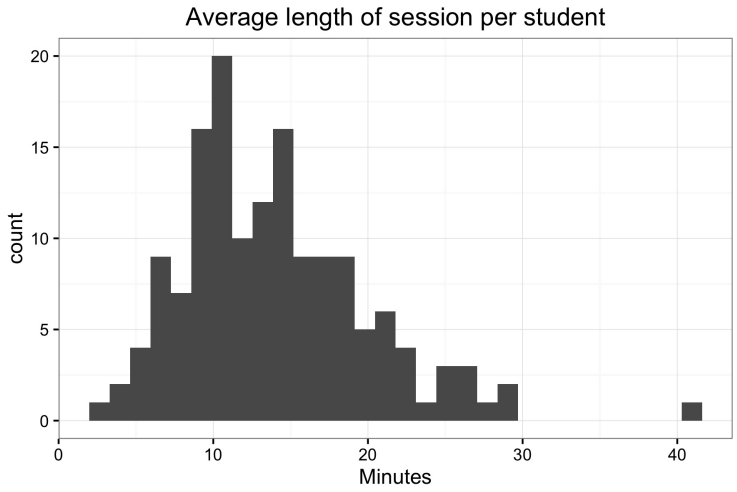

Figure 3.6: The average length of a session (each point represents a student).

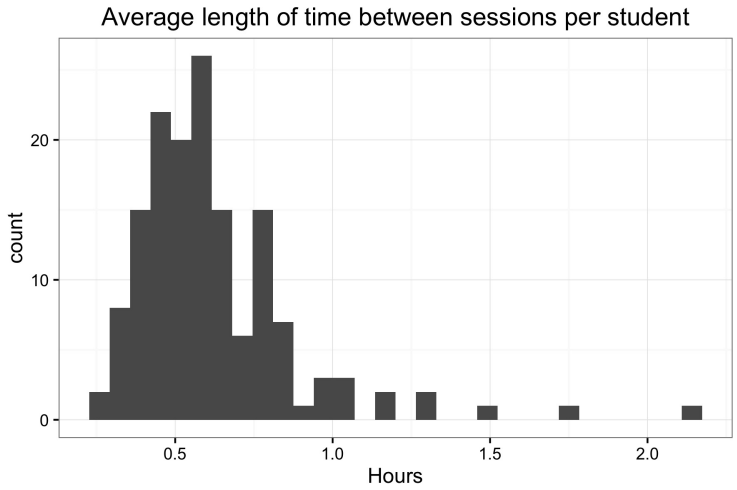

Figure 3.7: The average length of time between sessions (each point represents a student).

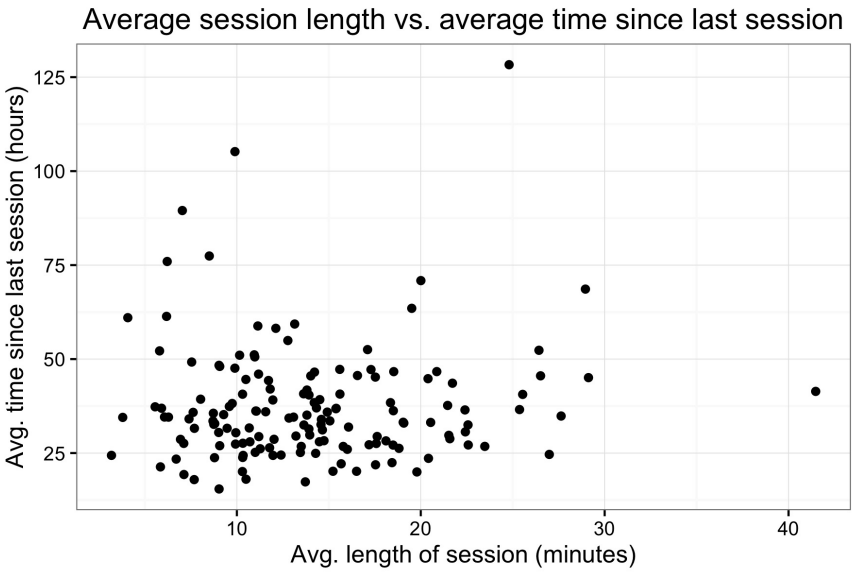

Figure 3.8: The average length of a session versus average length of time between sessions.

The median length of a session is 13.5 minutes, and the median elapsed time between sessions is 34.5 hours. There is not a correlation between session length and time elapsed between sessions ($r = 0.06, p = 0.46$).

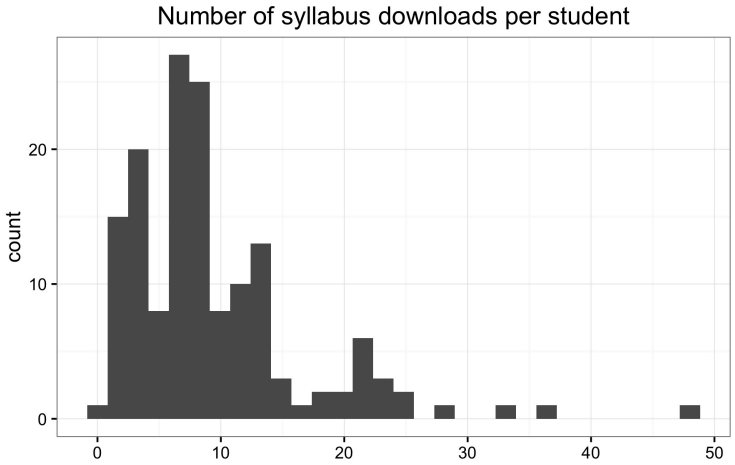

Figure 3.9: The number of syllabus downloads by student (each point represents a student).

The number of times a student downloads the course syllabus is highly right skewed (Figure 3.9), suggesting that students do not save the syllabus to their computer and refer to their local copy. (Even if the instructor edited the syllabus several times over the course

of the quarter, the right skew still suggests that some students are downloading the syllabus more than the number of times edited–or that most students are downloading fewer than the number of edits, which is unlikely. In this case, the instructor did not edit the syllabus after it was uploaded, which confirms that the right skew is severe.)

On average, a student in this course uses the CCLE site every day and a half for almost 15 minutes, clicking just over 7 links each time. Students that have more sessions tend to have fewer clicks per session, though the fact that they have more sessions means that they also have more overall clicks. Students that have more sessions do not necessarily have longer sessions ($r = -0.08, p = 0.34$).

3.2 Identifying Patterns in Student Usage of CCLE

To identify patterns in how students used the CCLE page, several variables measuring student activity were used in k-means clustering. The final variables used for clustering were the number of events, the number of sessions, the number of syllabus downloads, the number of non-syllabus file downloads, the number of times the announcement forum was opened, the number of announcement forum posts read, the number of times the discussion forum was opened, and the number of times a discussion post was made.

Because there are eight variables involved in clustering, visualizing the clusters on the original variables would require an eight-dimensional graph. A two-dimensional graph can be produced by transforming the data via principal component analysis (PCA), clustering using the data transformed by the first two principal components (PCs), and graphing the results. PCA will produce variables that are ordered by the proportion of data captured. Using the first two PCs to transform the data will allow for a two-dimensional visualization, but in order to preserve as much of the information for the clustering analysis as possible, the first two PCs need to capture as much variance in the data as possible. The inclusion of average session length and average time between sessions was considered, but they reduced PCA's ability to capture the variance in the first two principal components (PCs), so they were excluded from the analysis.

	PC1	PC2
With session time variables	42.5%	60.1%
Without session time variables	47.2%	66.8%

Table 3.12: Cumulative proportion of variance captured by the first two principal components in PCA, with and without session time variables.

Excluding the session time variables, the first two PCs (Table 3.13) show that the first PC describes general student activity, such that any event results in a decrease in PC1. The second PC describes more passive events (opening the announcement or discussion forum), but also more collaborative and interpersonal events (reading an announcement post or making a discussion post), such that both the passive and collaborative events result in a decrease in PC2.

	PC1	PC2
Number of events	-0.46	0.29
Number of sessions	-0.41	0.27
Number of syllabus downloads	-0.34	0.15
Number of announcement posts read	-0.38	-0.42
Number of times ann. forum opened	-0.18	-0.43
Number of times disc. forum opened	-0.38	-0.41
Number of discussion posts made	-0.33	-0.14
Number of non-syllabus downloads	-0.28	0.53

Table 3.13: First two principal components.

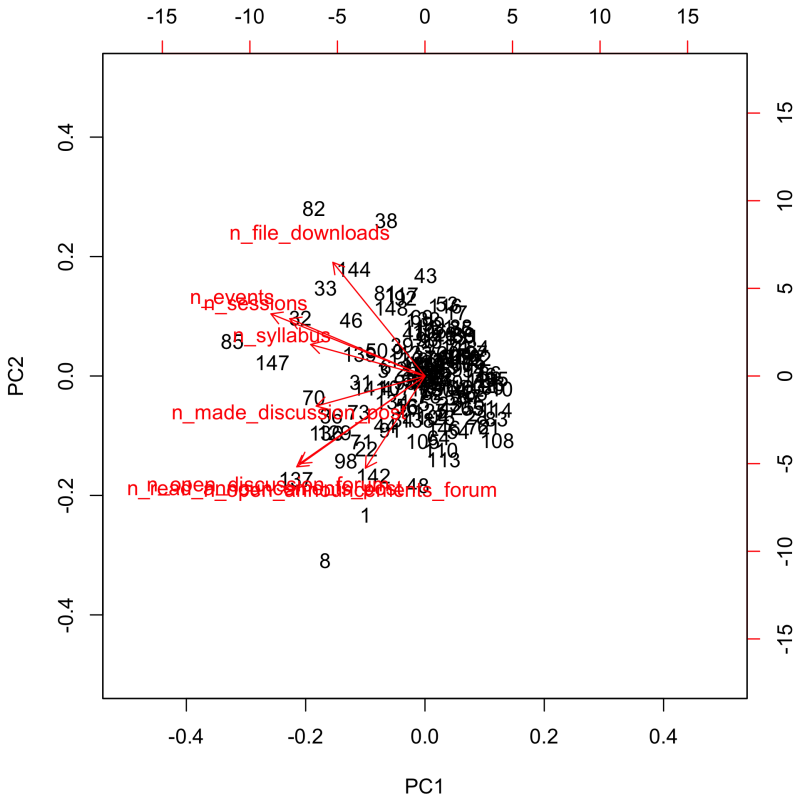

Figure 3.10: Biplot of the first two principal components.

The data were transformed by the first two PCs, resulting in a two-dimensional dataset. k-means clustering was performed on the transformed data, varying k from 2 to 10 to determine the best balance between reducing the sum of squared errors and maintaining a small number of clusters for interpretability. The "elbow method" was used to determine the k value that met this balance. $k = 3, 4, 5$ were chosen for further examination. (See Figures A.1 to A.3 for examples of elbow plots.) Although $k = 5$ had the lowest sum of squared errors, it was also the most unstable in its results. (Out of 102 attempts at clustering with $k = 5$, the most frequent result was only seen 20 times, and it did not have the smallest within sum of squares.) The clustering with $k = 3$ had the most stable results (out of 100 attempts, the most frequent result was seen 82 times), so it is used to identify patterns in student activity in CCLE.

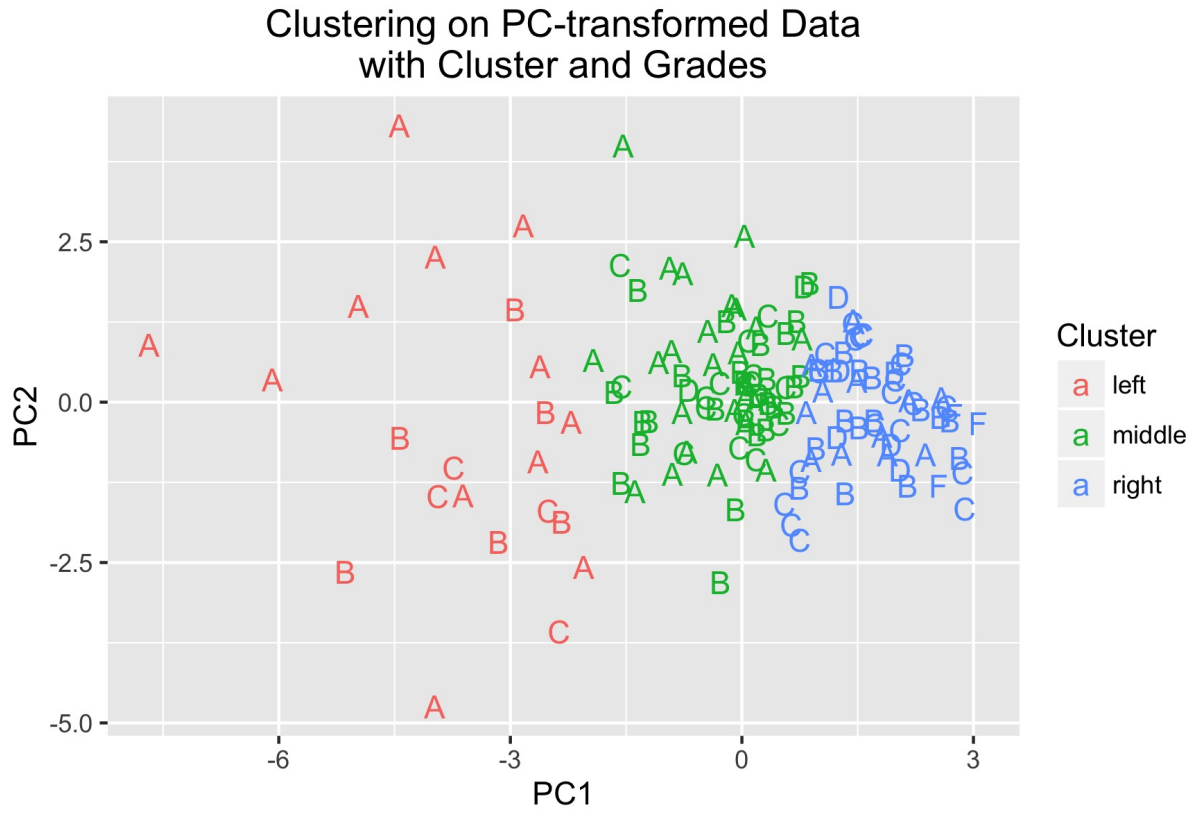

Figure 3.11: Students by k-means cluster and grade, graphed on first two PCs after PC transformation.

The three clusters are divided mostly along PC1. This is expected, given that PC1 captures the most variance in the data of all the PCs. The "leftmost" cluster–the most negative along PC1–contains no failing grades (Ds or Fs). The "middle" cluster has two Ds, and the "rightmost" cluster–the most positive along PC1, indicating a lack of general activity on CCLE–is associated with the greatest number of failing students. This suggests that, for this course, a higher level of activity is associated with higher grade performance. Figure 3.12 shows an association, that the more rightward the cluster–that is, the less general activity on CCLE–the greater proportion of failing students.

The reverse fanning effect in Figure 3.11 is expected, given that all the CCLE activity data used for clustering is count data, with a lower bound at 0.

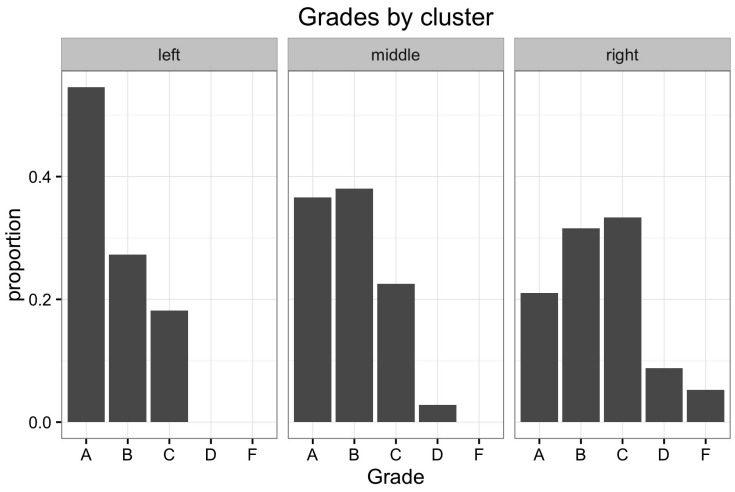

Figure 3.12: Full-letter grade breakdown by PCA-transformed k-means clustering.

The leftmost cluster, with no failing students, also shows an interesting division: there are no C grades when their value along PC2 is positive. Table 3.14 gives a detailed description of grades by cluster and whether the student has a positive or negative value along the PC2 dimension. While the middle cluster has more failing students with positive PC2 values, the rightmost cluster has far more failing students with negative PC2 values. In fact, in all three clusters, a negative PC2 value is associated with a downward shift in the grade distribution. An increase in the collaborative, participative events on CCLE are associated with poorer performance in the course, suggesting that looking for more help by reading announcement forum posts on CCLE or participating in the discussion forum is indicative of struggling in the class.

Cluster	PC2	A	B	C	D	F
Left	+	7	1	0	0	0
	-	5	5	4	0	0
Middle	+	17	14	10	2	0
	-	9	13	6	0	0
Right	+	6	7	9	2	0
	-	6	11	10	3	3

Table 3.14: Full-letter grade breakdown by PCA-transformed k-means clustering and whether the PC2 value is positive or negative.

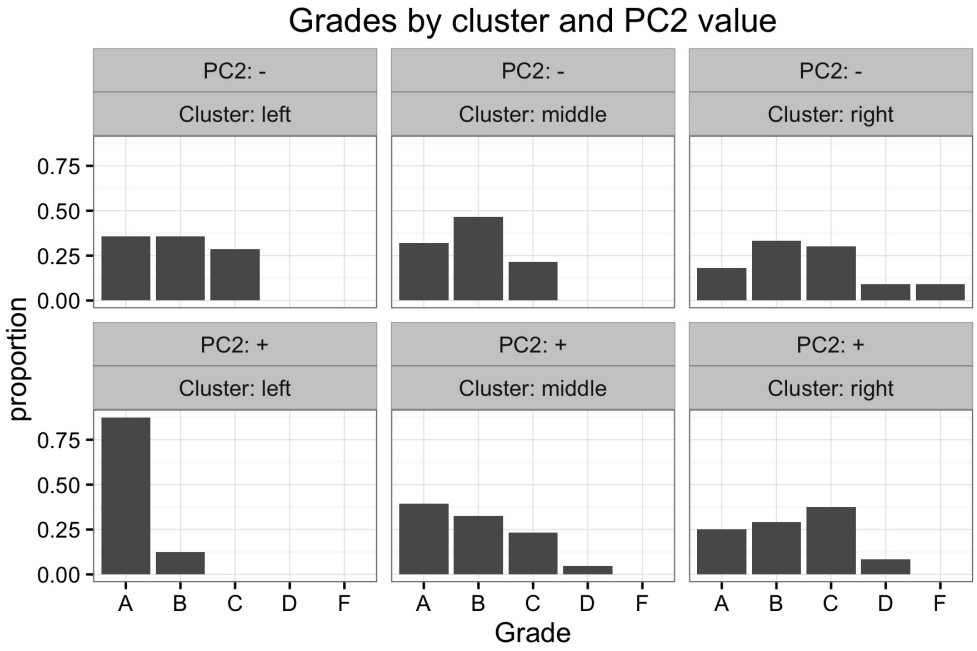

Figure 3.13: Full-letter grade breakdown by PCA-transformed k-means clustering and whether the PC2 value is positive or negative.

Clustering the PC-transformed data using k-means and the first two PCs shows that the clusters are separated are by general student activity, although Figure 3.11 shows that there is not much space between the three clusters. Nonetheless, the PC transformation shows that failing STATS 10 is associated with a low level of activity on CCLE. All of the variables have

a negative PC1 value, so a negative value for PC1 is associated with a higher level of general activity on CCLE. The variables with negative PC2 values relate to forum participation. Because a negative value in the PC2 dimension is associated with a poorer grade, it appears that students who seek collaborative assistance on CCLE may do so because they know that they are finding the course difficult.

CHAPTER 4

Predictive Modeling of Student Grades

The availability of CCLE data presents an opportunity for use in predictive modeling, with student grade as the outcome. The baseline model for comparison will be a model with only individual assignment grades as the input data, because these are the main data available to the instructor to assess how well a student is doing in the course. Including more data than is available to the instructor will only be useful if it improves upon his or her predictive ability from student assignment grades. This chapter compares predictive models with four possible formats for the grade outcome, as well as for three groups of input data sources. Finally, the change in predictions over time will be examined. Among all the comparisons, SVM regression using all available data, sorted into pass/fail grades using a cutoff value, is the best performing model.

These analyses will be performed using random forests (RF) and support vector machines (SVM). The SVM models created in this chapter use a radial basis function kernel to transform the data.

4.1 Model Building

In the following comparisons, baseline models will be created, with a standard set of input data that includes CCLE activity from the first five weeks of the quarter, individual assignment grades from the first five weeks, and selected demographic variables (sex, ethnicity, first term at UCLA, transfer status, residence status, Pell grant recipient status, and whether or not the student is the first in their family to complete a bachelor's degree). Other models can be built to predict student grade by varying the format of the outcome (half-

letter, full-letter, passed/not passed, percentage) or the input data sources (CCLE activity, demographics, individual assignment grades). The baseline models will be compared against models that vary the outcome variable and the input data source.

4.1.1 Baseline Model and Evaluation Criteria

The baseline model is built to achieve the best predictions a course instructor could make with the data available to him or her. In these models, classification and regression are performed only using individual assignment grades from the first five weeks of the term. (The regression predictions provide mean absolute error (MAE) for predicting students' final percentages in the class.) Models will be considered better than the baseline if the inclusion of CCLE activity and/or demographic data improve prediction results.

Outcome	Error (RF)	Error (SVM)
Full-letter	39.3%	22.0%
Pass/Fail	5.33%	3.33%
Final Percentage	5.635 (MAE)	4.227 (MAE)

Table 4.1: Baseline models with only individual assignment grades: error rates.

Of note is that the overall misclassification error, as seen in Table 4.1, is not a helpful measurement for this thesis's goal. Given that the goal of this thesis is to test the feasibility of an early warning system for students who are likely to fail the course, a failing grade should be deemed as a "positive" result; then the true positive (TP) rate for detecting failing grades measures that goal more effectively. The false positive (FP) rate is also important for assessing how many passing students would mistakenly be informed that they are in danger of failing the course. The best model will strike a balance between a high true positive rate and a low false positive rate.

In order to compare TP and FP rates across the different outcome variables, the classification results are converted to a pass/fail scheme using D+ and below as the cutoff grade for failing the course (see Table 2.2), and the regression results predicting final percentage

grades use 64 and below as the cutoff for failing the course. This method of converting classification and regression results to pass/fail grades will also be applied to all the non-baseline models to facilitate comparison among the models.

	RF		SVM	
Outcome	TP	FP	TP	FP
Full-letter	30%	5%	50%	0%
Pass/Fail	40%	1.43%	60%	0.7%
Final Percentage	70%	6.43%	100%	3.57%

Table 4.2: Baseline models with only individual assignment grades: true positive and false positive rates. (All results sorted into failing grades as a "positive" result.)

4.1.2 Comparison by Input Data Sources

One way to test for improvement over the baseline model is to add new data to see how additional information changes the predictive accuracy of the models. Because the highest true positive rates come from regression in both RF and SVM, the results in Table 4.3 likewise display the results for both RF and SVM in regression using the cutoff value for passing grades.

		RF			SVM	
Data sources	MAE	TP	FP	MAE	TP	FP
CCLE	10.494	0%	5%	7.515	0%	0%
CCLE, assignments	5.917	70%	3.57%	3.708	90%	3.57%
CCLE, assignments, demographics	5.454	80%	2.86%	2.873	100%	2.14%

Table 4.3: Error rates by input data sources for regression on final percent scores. (All results sorted into failing grades as a "positive" result.)

Adding more data decreases the MAE for RF and SVM regression. In addition, the CCLE activity-only models for both RF and SVM perform more poorly than the baseline

models, indicating that the individual assignment grade data have a stronger "signal"–better information for prediction–than the CCLE activity data alone has.

Using both individual assignment grades and CCLE activity data result in marginally worse predictions for RF compared to the baseline model, but improved results for SVM, as measured by MAE. Using all three possible data sources results results in a marginally smaller MAE for RF compared to the baseline, and continues to improve the MAE for SVM.

4.1.3 Comparison by Outcome Variable

Another issue in creating the predictive model is determining the format of the outcome variable. In the CEA data, half-letter grades are provided, including "P" and "NP" denoting passed and not passed, respectively. (See Table 3.1 for the distribution of half-letter grades.) These can be collapsed into full-letter grades by removing the "+" or "-" suffix from each grade, and assigning "P" to a passing grade (here, a "B") and "NP" to a failing grade (here, a "D"); or further collapsed into passed/not passed (or "pass/fail") grades by assigning As, Bs, Cs, and Ps to a passing grade, and Ds, Fs, and NPs to a failing grade (see Table 2.2 for a visual depiction of the relationships between outcome variables). The individual assignment grades for the course also contain a column for final percentage grade, offering a numeric possibility for the outcome. Thus classification–both binary and multi-class–and regression models can be created using the data.

All of the following models were built using the CCLE activity, individual assignments, and demographics data, as recommended by the previous section. CCLE activity and individual assignments from the first five weeks of the term only–not the full-term information–were used to train the models.

Outcome	Error (RF)	Error (SVM)
Half-letter	64%	37.3%
Full-letter	42%	13.3%
Pass/Fail	6%	2.67%
Final Percentage	5.843 (MAE)	2.873 (MAE)

Table 4.4: Overall error rates by different outcome variables.

The overall error rate decreases as the number of classes decreases from half-letter grades (15) to full-letter grades (5) to passed/not passed (2), which points to another reason the misclassification rate is not a suitable measure of error for this thesis (the first is that the goal focuses on detecting failing grades, rather than correct overall classification). It is disparate to compare misclassification rates across the different outcome variables because of the changing number of classes. In the half-letter grades, if the grades were distributed at random, 1/14 of the grades would be expected to be classified correctly; for full-letter grades, 1/5 would be expected correct; and for pass/fail grades, 1/2 would be expected correct. Decreasing the number of classes is expected to artificially decrease the overall misclassification rate.

Thus, using a cutoff value to predict passing/failing grades produces the best outcome variable if overall accuracy is desired. Because the evaluation criteria are maximizing the TP rate and minimizing the FP rate, the best outcome variable may be different. Collapsing each outcome variable to a pass/fail scheme allows the computation of TP and FP rates.

	RF		SVM	
Outcome	TP	FP	TP	FP
Half-letter	20%	1.43%	70%	0%
Full-letter	10%	1.43%	60%	0%
Pass/Fail	30%	1.43%	60%	0%
Final Percentage	70%	4.29%	100%	2.14%

Table 4.5: Error rates by different outcome variables. (All results sorted into failing grades as a "positive" result.)

The TP rate for predicting half-letter grades is better across both algorithms than the TP rate for predicting full-letter grades. RF for binary classification of passing/failing grades performs better than the other two RF classification methods, while SVM performs about the same as the other two SVM classification methods. None of the SVM classification models have false positives, meaning that no passing students were incorrectly classified as failing. Predicting final percent grades by regression and then using below 65% as a cutoff for failure is the best method in terms of identifying the most failing students, though its false positive rate is the highest among all the models in its algorithm.

SVM regression is the best performing model overall, as it identifies all failing students, though it incorrectly identifies several students as failing, even when they did not. These false positives will be discussed in the findings.

For the remainder of this thesis, this SVM regression model, with CCLE activity, individual assignments, and demographics data will be referred to as the "best model."

4.1.4 Measuring Student Progression over the School Term

To see how predictions change over time, the week-by-week prediction results of RF and SVM regression were compared. 11 models were created, one for each of the 11 weeks of the term (ten weeks of instruction and one week of final exams [17]). The input data for each model corresponds to the information available at that time: CCLE activity and individual

assignment grades were sliced so that only data up to the end of the week were used in the input dataset. (Because the demographic variables do not have a timestamp, they were included in every model.) Thus each regression model makes predictions using all available data up to the end of the week.

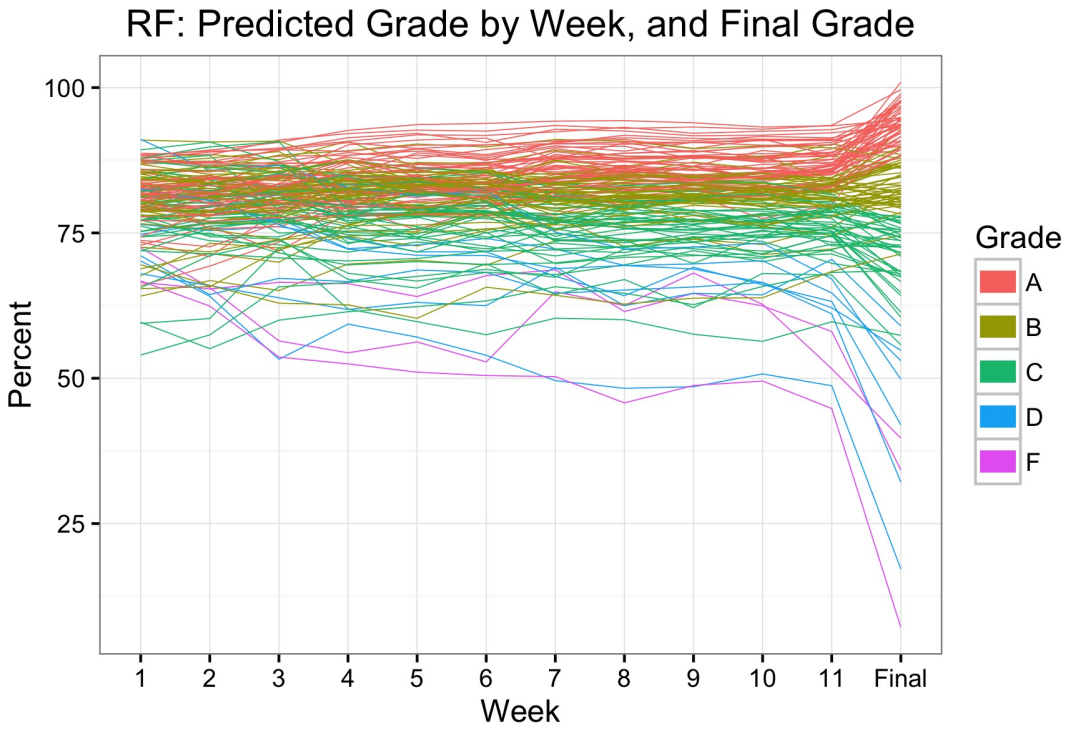

Figure 4.1: RF-predicted values of student grade using data available from weeks 1-11, compared against actual final grade.

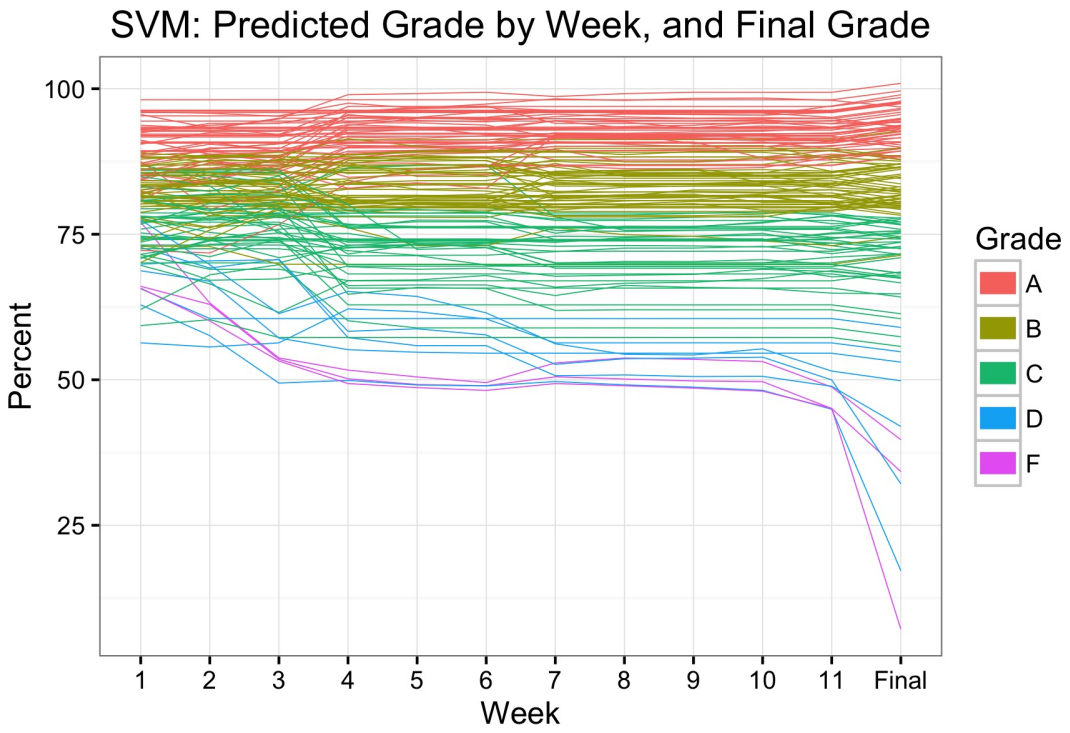

Figure 4.2: SVM-predicted values of student grade using data available from weeks 1-11, compared against actual final grade.

The SVM regression results are quite stable compared to the RF regression results, and they are also more accurate–the lower MAE can be seen in the similarity between the week 11 predictions and final percentages (with exceptions for large MAEs for failing grades). The RF prediction results from week 11 seem to shrink toward the middle, as evidenced by the fanning that occurs between the week 11 predictions and final percentages.

The average full-letter grades for SVM prediction show discernible trends. A-level grades are predicted to improve over the term, B-level grades stay constant, C-level grades drop slightly, and failing grades drop a greater amount.

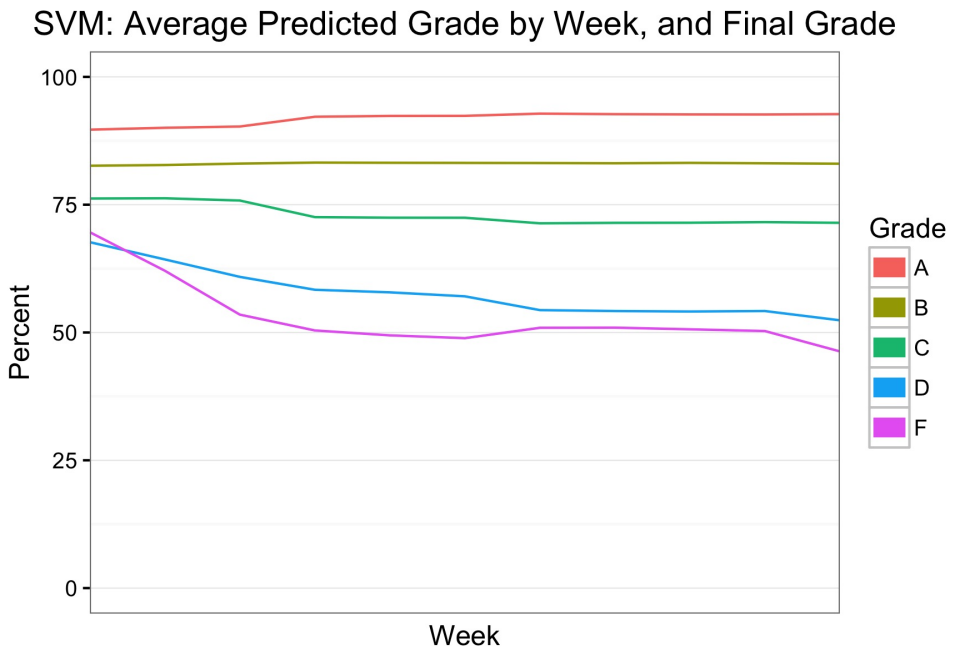

Figure 4.3: SVM-predicted values of average student grade using data available from weeks 1-11, compared against actual final grade.

4.2 Findings & Interpretations

4.2.1 Instability and Class Imbalance Issues in Random Forest Classification

Random forest results are variable by virtue of the fact that they "are a combination of tree predictors such that each tree depends on the values of a random [sampled] vector" [4]. Although "the Strong Law of Large Numbers shows that they always converge" [4], the convergence of one random forest need not be the same as the next; thus there is the potential for instability in the results. If many random forests all had similar results, then the model could be said to be relatively stable. This is shown not to be the case with the random forest models in this thesis. The results of the random forest models are not very stable, and running the same model several times–which changes the variables that are selected for each branch–affects the results. Table 4.6 counts the true positive rates for 100 random forest models for full-letter classification, using all data sources, and the same parameters each

time. The most frequent true positive rate is 0.2, but a rate of 0.3 occurs almost as often when *ntree* is 1000. This illustrates the issue of instability in the random forest models.

	True positive rate				
ntree	0	0.1	0.2	0.3	0.4
500	3	23	37	33	4
1000	1	22	48	29	0

Table 4.6: Frequency of true positive rates for 100 random forest models for full-letter classification.

While Table 4.6 shows that the most frequent true positive rate for these particular random forest models is 0.2, in each case, over 50% of the trees had a different result. Moreover, it is time consuming to create 100 random forests with 1000 trees each, especially if more data were to be added (such as from additional courses).

The instability in classification could be due to the fact that there is a high class imbalance (140 passing and 10 failing students). RF can "suffer from the curse of learning from an extremely imbalanced training dataset. As it is constructed to minimize the overall error rate, it will tend to focus more on the prediction accuracy of the majority class, which often results in poor accuracy for the minority class [8]. For this particular issue, adding more data would not help, unless the new data contain higher proportions of failing students. Up-weighting the cost of misclassifying the minority class or stratified sampling could resolve the class imbalance issue[8], and could be an avenue for future study with RF prediction.

The easiest method to address the class imbalance would be to prefer random forest regression over classification, as regression does not rely on the balance of cases to train the model. Table 4.2 shows that even in the baseline models, the results for random forest regression (after being separated into pass/fail grades) results in higher true positive rates of detecting failing students. Table 4.5 confirms this finding when using the best set of data sources.

Notably, SVM does not suffer from the same instability issues. In a fully separable dataset, whether that is before or after transformation to a higher dimension, "[t]he optimal hyperplane... is the unique one which separates the training data with a maximal margin" [9]. Even in the case where the training data are not fully separable, "the solution to the problem of constructing the soft margin classifier is unique and exists for any data set" [9]. Without the randomness inherent in constructing a random forest, SVM is not subject to variation in results.

4.2.2 Effectiveness of Random Forests versus Support Vector Machines

The best support vector machine predictions came from regression on final percent grade using CCLE activity, individual assignment, and demographic data. The best random forest predictions came from the same circumstances. (Due to randomness, the best RF predictions came from the model used to compare input data, rather than the model used to compare outcome variable though the input data, outcome variable, and parameters were the same.)

	RF		SVM	
	Fail (Predicted)	Pass (Predicted)	Fail (Predicted)	Pass (Predicted)
Fail (Actual)	8	2	10	0
Pass (Actual)	4	136	3	137

Table 4.7: Confusion matrices for best random forest and support vector machine models.

The SVM regression model is superior to the RF regression model both in detecting failing students and avoiding classifying passing students as failing. One likely reason for this is that SVM "maps the input vectors into some high dimensional feature space Z through some non-linear mapping chosen a priori" [9]–this is one of the distinguishing procedures in the SVM algorithm. The SVM models in this thesis all use the radial basis kernel for the non-linear mapping, and it seems that the data are highly separable in this space; however, this conjecture is difficult to affirm with a visualization because the input dataset has 24

variables. Projecting this down into two-dimensional space would likely lose a lot of the variation in the data.

4.2.3 Usefulness of the Data for Prediction

The comparison of the data inputs shows that CCLE activity data alone are too noisy to use as the sole data for grade prediction. In Figure 3.11, the final full-letter grades projected onto the first two PCs shows that the grades are close together. For example, around the coordinates (-3.75, -1.25), there are two Cs next to an A. A more exaggerated instance is near (-0.75, -1), where there is an A near a D and an F.

One intuitive hypothesis as to why the CCLE activity data alone are a noisy input source is that high levels of activity may be caused by students who do well in the course and are engaging with all the online material, or they may be caused by students who do poorly and are frequently opening resources to try to master the course. A similar dual explanation can be given for low levels of activity: they may be caused by students who do well and know the material, and thus do not need to view the resources often, or they may be associated with students who do poorly precisely because they do not review the material posted online. With the CCLE activity data alone, it is difficult to isolate the associations.

It makes sense, then, that even the baseline model with only individual assignment data performs better than the model with only CCLE activity data. (See Table 4.3.) The assignment grades are a direct measurement of student success in the course; in fact, they are the only official measures of student success, such that the final grade is typically a linear combination of all assignment grades. As seen in Tables 3.10 and 3.11, there are differences in distributions by demographic variables. These differences are distinct enough that both RF and SVM are able to pick up on them and reduce both MAE and the false positive rate over the baseline model.

The MAE for SVM with only assignment and demographic data is 3.363, with a TP rate of 100% and a FP rate of 2.86%, which is outperformed by the model with CCLE activity,

assignment, and demographic data. Thus while CCLE activity alone is a poor predictor of student success, its inclusion in a SVM regression model strengthens its predictive ability.

4.2.4 False Positives in Regression Modeling

The regression methods, whether RF or SVM, are the only methods to have false positives. This is because in the final percentage scores, there is an overlap of scores for C-level and D-level grades (that is, there is an overlap of scores between passing and failing grades). Thus even if the regression model were to have perfect accuracy, with a MAE of 0, using the simple rule of a cutoff value to determine passing and failing grades will produce errors. The errors will be false positives if the cutoff is set high to classify all failing grades correctly, or false negatives if the cutoff is set low to avoid misclassifying passing grades.

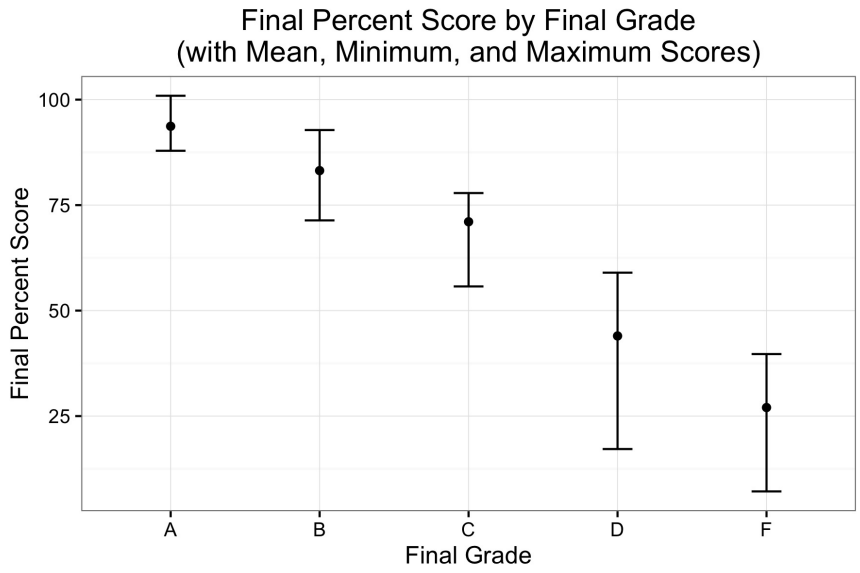

Figure 4.4: Distributions of final percentage scores by letter grade, with markers for minimum, maximum, and mean values.

The recommendation based on the results of this analysis is to set a cutoff value high enough to classify all failing grades correctly, achieving a true positive rate of 100%. The practical cost of false negatives is generally higher than the cost of false positives: it is worse to miss detecting and warning a student who is likely to fail, than to erroneously inform

a student that they are likely to fail when they are not. The cost could be reversed if a passing student who was mistakenly informed became so upset that they complained about the algorithm and warning procedure, but the results of the best SVM regression show that this is unlikely to happen.

The three passing grades that were misclassified are one C and two C-s, and as such are grades that are on the cusp of failing; these are not students who ended up with As or Bs where they would be shocked by a notification that they could fail this course. In fact, in Table 4.8, the difference between the predicted and actual final percent scores for the three misclassified grades is about 1.5 points, less than the overall MAE for the model. Moreover, while their actual final scores are above a rough cutoff threshold of 55, there is one NP grade with an actual final score of 58.99, which causes the overlap between passing and failing grades.

Predicted	Actual	Final Grade
48.65	7.15	F
49.10	17.19	NP
49.19	34.23	F
50.48	39.70	F
54.73	53.03	D
55.86	41.98	NP
57.26	55.73	C-
58.76	32.12	NP
58.91	57.38	C-
60.52	58.99	NP
61.69	54.81	D+
62.87	61.34	C
64.32	49.83	D

Table 4.8: Final grades associated with predicted (via best SVM model) and actual final percentage scores.

4.2.5 Variable Contribution in SVM Regression

The lack of interpretability and the ability to explain variable contribution is one disadvantage of SVM regression. While weights can be extracted from the SVM regression, they are difficult to interpret as they describe the separating hyperplane in radial basis function kernel space, which is infinite-dimensional [2].

CHAPTER 5

Remarks on Generalizability

Because this thesis functions as a pilot study for a UCLA-wide implementation of system to alert students and instructors of failing students, the generalizability of these findings is a primary concern.

In order for the findings in this thesis to generalize well, the circumstances between STATS 10 and the UCLA undergraduate population must be similar. This thesis used χ^2 tests on the demographic variables sex, residence status, and ethnicity to show that there are demographic similarities between the two populations. The commonalities must extend beyond basic demographics, though; for this thesis to generalize well, STATS 10 must be similar to the UCLA undergraduate population beyond ways measured by Registrar's Office demographic variables.

For one, STATS 10 must be similar content-wise to all other undergraduate courses. STATS 10 differs from a humanities course, where there may not be regular homework assignments, and which may cover much less mathematical material. A life science course may require more memorization than interpretation of findings (as with the results of a statistical test), and a different physical science course may require much more math than STATS 10. Upper division courses differ in scope and difficulty from lower division courses.

Moreover, STATS 10 must be similar in CCLE usage to all other undergraduate courses. For this particular class, the instructor utilized file uploads, announcement forum posts, and an online syllabus; he did not use online quizzes, and homework was submitted in person by hard copy, rather than digitally using a CCLE submission portal. Other courses may utilize both of those features. Some courses choose not to use CCLE at all, and instructors may

use their own website for lecture slides and handouts, or they may not maintain any online resource for the class at all.

Knowing that there are dissimilarities between STATS 10 and other UCLA undergraduate courses, tests for generalization are proposed.

5.1 Proposed Steps

This thesis focused on answering the preliminary question of whether failing students could be detected by available data at UCLA. As such, it did not focus on achieving the best possible model, nor did it use a large amount of training data. These limitations should be addressed in future studies to work toward predictive models that accurately classify failing students across a wide array of undergraduate courses. The following are some recommendations for how to adopt the results of this thesis for the UCLA undergraduate population.

The first major step will be to test the best SVM model on new data to see whether the model generalizes. Testing for generalizability should occur in several steps. Other recent courses by Professor Gould should be used first, as they are controlled for the instructor, and will thus have similarities to this STATS 10 course in terms of course content and the way CCLE is used. After this, there are many other populations of courses that can be used for testing: other courses in the statistics department should have similar course content, so the model has a good opportunity to generalize well. A humanities or social science course where the instructor uses CCLE comparably is also an area where the model may generalize well. A difficult test of generalization would be a humanities or social science course where the instructor uses CCLE sparsely. Under this scheme, this STATS 10 course is training data, and other recent courses are testing data.

If the SVM regression model is able to achieve a high true positive and low false positive rate in predicting failing grades for other courses, it can be considered successful. Success will determined by whether the model meets thresholds for a high true positive rate and a low false positive rate. The desired thresholds should be determined by project stakeholders beforehand. If the SVM regression model performs more poorly than the thresholds, then

new models can be created, or other aspects of the project (such as input data sources and variables) can be examined.

It should be anticipated that the SVM regression model will not produce as effective results on the new courses that constitute testing data; "[t]ypically, the training error... will be less than the true error... because the same data is being used to fit the method and assess its error" [10]. That is, "[a] fitting method typically adapts to the training data, and hence the apparent or training error... will be an overly optimistic estimate of the generalization error" [10]. This thesis is therefore optimistic, having presented only the training error from the predictions for the training data.

A second step would be to train SVM regression models on an entirely new set of data. This step would be especially helpful if the best SVM regression model in this thesis fails to meet the desired thresholds. The new training data could simply be the testing data just described. Alternatively, the training data could come from historical data on specific course-instructor pairings where detecting failing students is desired, and the testing data could be the most recent offering of the course by the instructor. For example, an instructor's three most recent STATS 10 classes could be collected for data, along with an instructor's three most recent CHIN 1 offerings. The two earlier offerings of each class would be the training data for a new model, and the most recent offerings of each class would be the testing data. This would allow new models to be fit not only on data from multiple courses, but also from different terms.

This approach of using classes over multiple terms as training data leverages UCLA's collections of CCLE and demographic data to provide more training data. While the traditional concern of using time series data is that the observations are not independent of each other, in this case, each observation in the dataset will be a student enrolled in a historical class. Students may appear more than once, if two or more courses in which they were concurrently enrolled are selected as training data, or if courses they were enrolled in during different quarters are selected. The probability of the same student appearing multiple times increases as more courses are used for training data. Thus while the observations will not necessarily be independent, the data are not time series data.

If a student does appear multiple times in the data, the decision to keep the multiple instances or to keep only one may be made according to domain knowledge or heuristics. If it is known that the way students interact with different courses is distinct enough that each appearance of the same student in the data can be considered almost independent of the others, then all such observations could be kept (or vice versa). A heuristic approach would be to use cross-validation on the models with and without the multiple appearances to see which performs better in terms of validation error.

In creating new models from new training data, it may be advantageous to build several models, separating courses according to their general field of study, or by the instructor's CCLE usage, or by availability of data (some instructors may not keep an easily accessible record of individual assignment grades). This would allow courses to be grouped by similarity, and the resulting models would have greater predictive ability.

During model building for this thesis, the cost parameters (`gamma` and `cost` in the `svm()` function from the `e1071` package in `R`) were left at their default values. For better model fitting, a grid search over those parameters is recommended [11].

5.1.1 Working with New Variables

Variables for this thesis were largely chosen based on ease of extraction from the data, as well as interpretability. Some of the metadata from the CCLE activity log were easily discernible, such as the timestamp or if a student downloaded a file by viewing a resource. (Table A.1 contains a sample row of the log where these entries can be seen.) Even so, not all of these variables were used in the analysis–some were excluded because they reduced the amount of variation captured by the first two principal components in PCA.

There are many potential new variables to identify from the CCLE log. First, the STATS 10 course used as the training data for this thesis does not utilize certain CCLE modules, such as quizzes or online submissions for assignments. Other courses that use these modules in CCLE will store the relevant student interaction as metadata in the log file, and these can be identified and cleaned as new variables for training data. Second, there may be variables

to extract even from the current STATS 10 log by those more familiar with the metadata and how they are collected. Existing variables could be transformed or combined to create new variables.

Other variables could be added from the demographic data, as well. The original demographic data came with many more variables than were used for the predictive modeling.

New variables can be added to the training data for new models, but this is not without its drawbacks: "failure to discard irrelevant features [variables] (e.g. noise, outliers, redundant features) will affect the system performance which includes classification accuracy... the implicit regularization achieved by feature pruning typically increases the generalization ability of classifiers" [14]. Generalization of the model can be tested using cross-validation: a way to check which new variables should be added to the model would be to create several different models with different sets of variables, and to check the validation error using 5-fold or 10-fold cross-validation.

5.2 Feasibility

It appears very feasible to extend the methods and results of this thesis to a cross-campus predictive system. The various data (CCLE, demographics) are already being collected and only need to be accessed. If the files are of a reasonable size, they can be processed with the code written for this thesis.

Assignment grades will be more time-intensive to process. If the data are stored in CCLE or another online system, they should be easily retrievable and processable. On the other hand, if the instructors keep a record of assignment grades in a local file, such as an Excel spreadsheet, these will need to be pre-processed so that they are all uniformly formatted and easily read by a statistical package.

This points to a concern about the feasibility of scaling up this predictive model, which is the time and computing ability it will take to train models on large amounts of data. Human time and effort will be expended to format the data manually for analysis, as in

the case of assignment grades in local files. Computing time and power will also need to be considered when the amount of training data increases. This study was conducted using `R`, which by default stores its objects in memory [22]; if the datasets are too large, then other solutions to store and process the data will need to be sought. Also, `R` is not a language written for processing speed [23], as evidenced by additional packages that exist solely to speed its performance [21]. Other methods in `R` or other statistical packages altogether may need to be employed to handle the increased demand for computing resources.

Still, a well-generalized model could be found using a smaller scale of data, which would allow the use of similar machines and scripts as this thesis. This would alleviate the need for significant changes to the model creation computing process. Moving from smaller to larger training datasets should indicate whether an increase in computing resources is necessary.

CHAPTER 6

Conclusion

Using data from CCLE, demographic information, and students' assignment grades, several random forest and support vector machine models were created. These models were compared against each other to see which could achieve the best classification of failing grades as a balance of true and false positive rates, as well as against baseline models that only used student assignment grades to simulate best predictions according to the data easily available to an instructor. The models differed on their input data sources (CCLE only; CCLE and assignment data; and CCLE, assignment, and demographics data), as well as on their outcome variable (half-letter grades, full-letter grades, pass/fail, and final percent grades). Despite the differences in outcome variables, the results could be compared by employing a cutoff method to sort half-letter and full-letter grades into pass/fail grades, and to sort predictions of final percent grades into pass/fail grades.

Among these models, the best model was an SVM regression model whose predictions of final percent grades were classified into pass/fail grades according to a cutoff value.

This model correctly classified all failing students, while misclassifying three passing students. These three students received a C and two C-s, and these three students had final percentage scores of between 55% and 62%. The misclassification is thus not severe. Additionally, as the two C-s had lower final percentage scores than a "not passed" grade. Because of this, false positives are difficult to avoid in the regression models using a cutoff score for failing grades.

To test whether the model generalizes well to other classes, it is recommended that it be used to predict grades using other classes for input data. Starting with other classes by the same instructor, and gradually incorporating more dissimilar classes, the tests should

indicate whether the best model from this thesis works well for other classes at UCLA. Other models can be created that use other classes in other quarters as input data, as well as using new variables identified from the CCLE log or from demographic information. With more data, these models may do a better job of generalizing to UCLA undergraduate courses.

As it is, this thesis demonstrates that it is possible to detect failing students in a course part-way through the term. If this model had been implemented for the STATS 10 course it draws data from, all failing students could have been notified via CCLE or the course instructor after the first five weeks of the term that they were in danger of failing the course. The major caveat is that this thesis only examines the training error as evidence for this claim. If the model or approach generalizes across the offerings of UCLA undergraduate courses, there is great potential for the ability to intervene as early as halfway through the quarter with students who are at risk of failing the course, with the goal that an early warning and extra attention will assist these students become proficient in the course content.

APPENDIX A

Appendix

A.1 Diagnostics for k-means Clustering

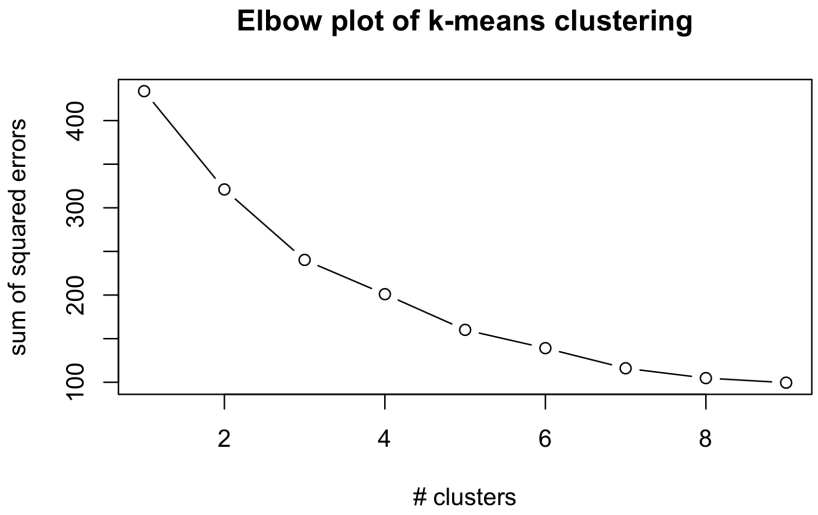

Figure A.1: Elbow plot of sum of squared errors by number of clusters (k); example 1.

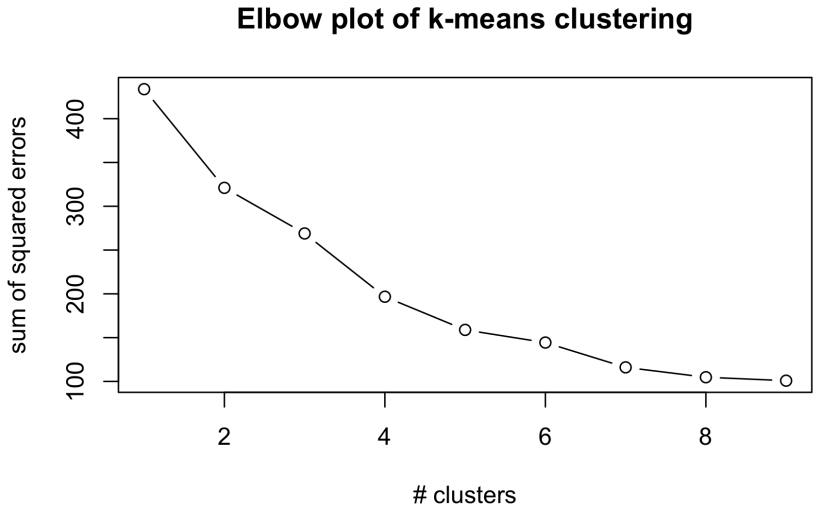

Figure A.2: Elbow plot of sum of squared errors by number of clusters (k); example 2.

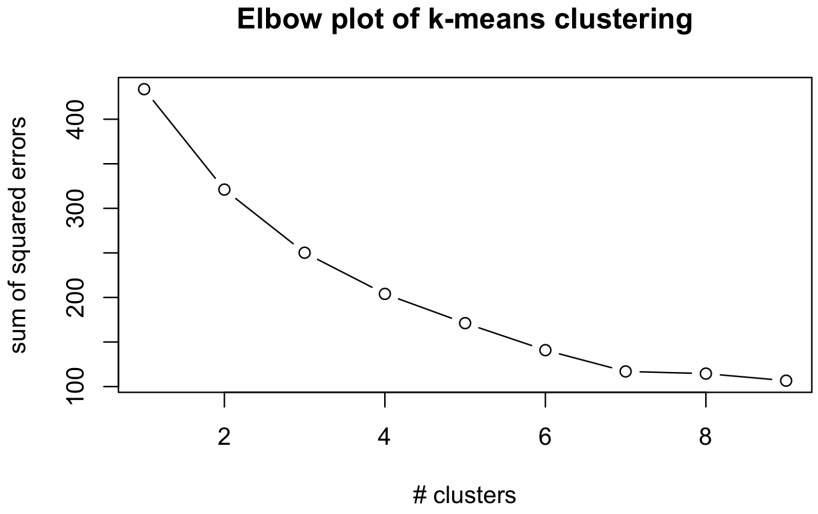

Figure A.3: Elbow plot of sum of squared errors by number of clusters (k); example 3.

A.2 Sample Row of CCLE Log

Column	Entry
id	177538596
eventname	\mod_resource\event\course_module_viewed
component	mod_resource
action	viewed
target	course_module
objecttable	resource
objectid	582052
crud	r
edulevel	2
contextid	1223394
contextlevel	70
contextinstanceid	996303
userid	45086
courseid	32961
relateduserid	NULL
anonymous	0
other	N;
timecreated	1454275706
origin	web
ip	[masked]
realuserid	NULL

Table A.1: Sample row from CCLE log, with masked IP address.

Bibliography

[1] UCLA Undergraduate Admission. Quick Facts about UCLA, n.d. URL http://www.admission.ucla.edu/campusprofile.htm. Accessed on May 18, 2017.

[2] Matthew Bernstein. The radial basis function kernel, 2017. URL http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/svms/RBFKernel.pdf. Accessed on June 1, 2017.

[3] Blackboard. What We Do, n.d. URL http://www.blackboard.com/about-us/what-we-do.aspx. Accessed on April 25, 2017.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Canvas. Canvas, n.d. URL https://www.canvaslms.com/. Accessed on April 25, 2017.

[6] UCLA CCLE. About CCLE, n.d. URL https://ccle.ucla.edu/course/view/aboutccle?section=0. Accessed on April 25, 2017.

[7] Pew Research Center. College students and technology, 2011. URL http://www.pewinternet.org/2011/07/19/college-students-and-technology/. Accessed on April 25, 2017.

[8] Chao Chen, Andy Liaw, and Leo Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110, 2004.

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20 (3):273–297, 1995.

[10] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[11] David Meyer and FH Technikum Wien. Support vector machines. *R News*, 1(3):23–26, 2001.

[12] Moodle. About Moodle, 2016. URL https://docs.moodle.org/32/en/About_Moodle. Accessed on April 25, 2017.

[13] Moodle. Moodle Documentation, 2016. URL https://docs.moodle.org/32/en/Main_page. Accessed on April 25, 2017.

[14] Minh Hoai Nguyen and Fernando De la Torre. Optimal feature selection for support vector machines. *Pattern recognition*, 43(3):584–591, 2010.

[15] UCLA Registrar's Office. General Education (GE) Courses Master List, n.d.. URL https://sa.ucla.edu/ro/Public/SOC/Search/GECoursesMasterList. Accessed on May 18, 2017.

[16] UCLA Registrar's Office. Grades, n.d.. URL http://catalog.registrar.ucla.edu/ucla-cat2016-58.html. Accessed on May 18, 2017.

[17] UCLA Registrar's Office. Academic & Administrative Calendar 2015-2016, n.d.. URL http://www.registrar.ucla.edu/Portals/50/Documents/calendar-archive/academiccalendar15-16.pdf. Accessed on May 30, 2017.

[18] Pearson. Pearson student mobile device survey, 2015. URL http://www.pearsoned.com/wp-content/uploads/2015-Pearson-Student-Mobile-Device-Survey-College.pdf. Accessed on April 25, 2017.

[19] UCLA Academic Planning and Budget. Enrollment demographics, Fall 2016, n.d.. URL http://www.aim.ucla.edu/tables/enrollment_demographics_fall.aspx. Accessed on May 18, 2017.

[20] UCLA Academic Planning and Budget. UCLA 2015-16 Undergraduate Profile, n.d.. URL http://www.aim.ucla.edu/pdf/UGProfile15-16.pdf. Accessed on May 18, 2017.

[21] R-core. Package 'parallel', 2016. URL https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf. Accessed on June 1, 2017.

[22] Bill Venables. Stored object caches for r, 2013. URL https://cran.r-project.org/web/packages/SOAR/vignettes/SOAR.pdf. Accessed on June 1, 2017.

[23] Hadley Wickham. *Advanced R*. CRC Press, 2014. Accessed via http://adv-r.had.co.nz/Performance.html on June 1, 2017.