Data and Programs for the paper:

***Is Big Data Better? LMS Behavioral Trace Data And Predictive Analytic Performance in Postsecondary Education***

Kelli A. Bird, Benjamin L. Castleman, Yifeng Song & Renzhe Yu

August 17th, 2022

**Data Files for the Analysis**

We have access to a de-anonymized copy of the Virginia Community College System (VCCS) administrative data starting from Fall 2000 through a Master Data Use Agreement with VCCS. Those data contain detailed information regarding each student’s demographic and pre-college background, program of study, courses taken, GPA, financial aid received, VCCS degree earned, and his/her enrollment/graduation records in other institutions (National Student Clearinghouse data).

Through the Master Data Use Agreement with VCCS, we also have access to a de-anonymized copy of the VCCS Learning Management System (LMS) data starting from Spring 2019 (stored in Canvas). Those data contain student x course x section level clickstream behaviors with timestamps recorded by Canvas throughout each semester, along with the online student activities in assignment submissions and discussion forum participations.

Due to the nature of the student-level VCCS data, we will not be able to make a public repository to share the data files.

**Program Code Files for the Analysis**

The codebase consists of three directories as follows. It is necessary to follow the order of those directories to run the scripts.

**01\_data\_cleaning\_and\_construction**

* The scripts under this directory perform the entire process of cleaning both VCCS Administrative and LMS data, as well as constructing the predictors used for building predictive models.
* Table 1, Table 4, Appendix Table A1, Appendix Table A5 in the paper are generated by the scripts in this section.
* Run the scripts following the order outlined below.

1. **create\_merged\_class\_and\_gpa.do**:

* Merge all VCCS Class and GPA files, so that the relevant term-level course-taking and GPA data could be easily retrieved in subsequent steps of building predictive models.

1. **cum\_gpa\_by\_term.ipynb**:

* Calculate the aggregated cumulative GPA value by student x term for all students included in the study sample.

1. **clean\_gpa2212.do**:

* This script cleans up the cumulative GPA data in the most recent term, which doesn't show up in the merged historical GPA file.

1. **find\_dual.do**:

* This script identifies the dual enrollment status at the student x term level within 2019 Spring - 2021 Spring.

1. **impute\_P\_plus\_and\_minus\_grades.do**:

* This script imputes the P+/P- grades assigned during the Spring 2020 term (when COVID outbreaks) following the percentage of A/B/C/D/F grades assigned in that term. The imputed grades will be used when constructing predictors related course grades

1. **calc\_avg\_grades.do**:

* The script computes the average grades of each course x term during the past five years, which will be used to construct course-specific predictors.

1. **create\_merged\_class\_data\_sample.do**:

* This script identifies all VCCS courses that are graded using A/B/C/D/F scale, which are potential samples to be included in predictive modeling.

1. **preprocessing\_LMS\_data.ipynb**:

* This script performs initial preprocessing of the LMS data:
* Merge individual school-level files
* Join them with the VCCS admin data
* Drop the course sections which show up in Canvas data but not in admin data
* Create the current and historical LMS data respectively

1. **further\_clean\_LMS\_measures.do**:

* This script further cleans the current LMS data by handling the missing values within the assignments and discussion forums measures.

1. **further\_clean\_historical\_LMS\_measures.do**:

* This script further cleans the historical LMS data by handling the missing values within the assignments and discussion forums measures.

1. **identify\_admin\_sample\_step1.do**:

* This script merges the VCCS admin data with the LMS data, and identifies the student x term observations to be included in the study sample for non-first-term observations.

1. **cum\_gpa\_by\_term\_2212.ipynb**:

* The raw VCCS data report the student GPA by student x college x term. This script will calculate the aggregated cumulative GPA value for the most recent term (Spring 2021), which will be used in the subsequent steps of constructing the predictors of the training/validation sample for the course performance prediction models.

1. **create\_cum\_gpa\_by\_term\_long\_format.do**:

* This script creates the student x term cumulative GPA values and then convert the data into the long format, which will be used to construct term-specific predictors.

1. **identify\_admin\_sample\_step2.do**:

* This script continues identify\_admin\_sample\_step1.do by merging in additional information to identify the student x term x college x course x section observations to be included in the study sample for non-first-term observations.

1. **create\_part1.do**:

* This script creates certain course-specific predictors for non-first-term observations.

1. **create\_sample.do**:

* This script creates the data set that contains the information of all prior courses taken, which will be used to construct predictors related to prior courses later on.

1. **preprocessing\_additional.do**:

* Additional processing for the VCCS Class files, so that the demographic and non-course-specific predictors can be constructed for non-first-term observations. Those pre-constructed predictors will be retrieved in subsequent steps when building the clean training and validation sets for course performance prediction models.

1. **find\_race.do**:

* Merge the VCCS Student files, extract demographic information of each student in the non-first-term observations, which will be used as predictors in building course performance prediction models.

1. **additional\_term\_specific\_predictors.do**:

* This script constructs additional term-specific predictors including share of 200-level, online, evening courses, and also identify the concurrent courses of each course x section in the study sample for all non-first-term observations.

1. **ever\_dually\_enrolled.do**:

* This script identifies whether each student has ever been dually enrolled up to the target term.

1. **cip\_and\_degreetype.do**:

* This script generates the 2-digit CIP and degree level of each student x term, for all non-first-term observations.

1. **identify\_instructors.do**:

* This script identifies all VCCS instructors who have taught the courses included in the study sample of all non-first-term observations.

1. **create\_instructor\_related\_predictors.do**:

* This script creates all instructor-related predictors, including full-time status, indicator for tenure, the average grade assigned to the course by each instructor, for all non-first-term observations.

1. **find\_prior\_terms\_gpa\_and\_enrl\_intensity.ipynb**:

* Reorganize the term-specific GPA and enrollment intensity data into dictionaries that map student x term to the lists of prior term GPA and term enrollment values, so that they can be processed later on to find out the trendline (slope) of term GPA and term enrollment intensity as predictors for grade prediction models.

1. **create\_term\_gpa\_and\_enrl\_intensity\_predictors.py**:

* For each student x term, calculate the trendline (slope) predictors corresponding to prior term GPA and prior term enrollment intensity, which will be used in constructing course performance prediction models.

1. **trend\_predictors.ipynb**:

* This script merges all individual files of trend predictors generated by "create\_term\_gpa\_and\_enrl\_intensity\_predictors.py" into one single table.

1. **updated\_online\_ind.do**:

* This script creates the indicators for online and in-person courses in the study sample for non-first-term observations.

1. **create\_course\_specific\_predictors.ipynb + parallelized\_prereq.py + parallelized\_repeat.py**:

* First open the script“**create\_course\_specific\_predictors.ipynb”**: the comments within this Jupyter Notebook shows when to run “**parallelized\_prereq.py**” and “**parallelized\_repeat.py**”.
* This script creates all course-specific predictors for the study sample of all non-first-term observations.

1. **create\_term\_specific\_predictors.ipynb**:

* This script merges all term-specific predictors constructed in previous steps together, for all non-first-term observations.

1. **create\_subject\_specific\_predictors.ipynb**:

* This script creates the subject-specific predictors: interaction between the subject the target course belongs to and the subjects the previously taken courses belong to.

1. **create\_concurrent\_admin\_predictors.ipynb**:

* This script creates all concurrent admin predictors for all non-first-term observations, and then append them to the lists of term-specific and course-specific predictors, respectively.

1. **standardize\_early\_clickstream\_predictors.ipynb + parallelized\_standardization\_of\_early\_clickstream.ipynb + add\_credit\_hours.do + create\_concurrent\_clickstream\_predictors.ipynb + finalize\_concurrent\_LMS\_measures.do**:

* First open the script “**standardize\_early\_clickstream\_predictors.ipynb**”: the comments within this Jupyter Notebook shows when to run “**parallelized\_standardization\_of\_early\_clickstream.ipynb**”, “**add\_credit\_hours.do**”, “**create\_concurrent\_clickstream\_predictors.ipynb**”, “**finalize\_concurrent\_LMS\_measures.do**”, respectively.
* This script creates all of the standardized and cleaned early-term clickstream predictors for all non-first-term observations.

1. **standardize\_full\_clickstream\_predictors.ipynb + parallelized\_standardization\_of\_full\_clickstream.py + parallelized\_historical\_1.py + parallelized\_historical\_2.py**:

* First open the script “**standardize\_full\_clickstream\_predictors.ipynb**”: the comments within this Jupyter Notebook shows when to run “**parallelized\_standardization\_of\_full\_clickstream.py**”, “**parallelized\_historical\_1.py**”, “**parallelized\_historical\_2.py**”, respectively.
* This script creates all of the standardized and cleaned historical full-term & early-term clickstream predictors.

1. **identify\_admin\_sample\_firstterm.do**:

* This script merges the VCCS admin data with the LMS data, and identifies the student x term observations to be included in the study sample for first-term observations.

1. **create\_part1\_firstterm.do**:

* This script creates certain course-specific predictors for first-term observations.

1. **processing\_additional\_firstterm.do**:

* Additional processing for the VCCS Class files, so that the demographic and non-course-specific predictors can be constructed for first-term observations. Those pre-constructed predictors will be retrieved in subsequent steps when building the clean training and validation sets for course performance prediction models.

1. **find\_race\_firstterm.do**:

* Merge the VCCS Student files, extract demographic information of each student in the first-term observations, which will be used as predictors in building course performance prediction models.

1. **additional\_term\_specific\_predictors\_firstterm.do**:

* This script constructs additional term-specific predictors including share of 200-level, online, evening courses, and also identify the concurrent courses of each course x section in the study sample for all first-term observations.

1. **cip\_and\_degreetype\_firstterm.do**:

* This script generates the 2-digit CIP and degree level of each student x term, for all first-term observations.

1. **identify\_instructors\_firstterm.do**:

* This script identifies all VCCS instructors who have taught the courses included in the study sample of all first-term observations.

1. **create\_instructor\_related\_predictors\_firstterm.do**:

* This script creates all instructor-related predictors, including full-time status, indicator for tenure, the average grade assigned to the course by each instructor, for all first-term observations.

1. **updated\_online\_ind\_firstterm.do**:

* This script creates the indicators for online and in-person courses in the study sample for first-term observations.

1. **create\_term\_specific\_predictors\_firstterm.ipynb**:

* This script merges all term-specific predictors constructed in previous steps together, for all first-term observations.

1. **create\_course\_specific\_predictors\_firstterm.ipynb**:

* This script creates all course-specific predictors for the study sample of all first-term observations.

1. **create\_concurrent\_admin\_predictors\_firstterm.ipynb**:

* This script creates all concurrent admin predictors for all first-term observations, and then append them to the lists of term-specific and course-specific predictors, respectively.

1. **standardize\_early\_clickstream\_predictors\_firstterm.ipynb**:

* This script creates all of the standardized and cleaned early-term clickstream predictors for all first-term observations.

1. **identify\_full\_vccs\_population.do**:

* This script extracts the key admin measures of all student x term x college x course x section in the VCCS population.

1. **summary\_stats\_1.ipynb**:

* Generate the summary statistics of admin data for the whole study sample.

1. **summary\_stats\_2.ipynb**:

* Generate the summary statistics of LMS measures for the whole study sample.

1. **summary\_stats\_population.ipynb**:

* Generate the summary statistics of admin data for the entire VCCS population.

**02\_predictive\_modeling\_for\_whole\_sample**

* The scripts under this directory fit the random forests models using the whole study sample, and generate the model performance evaluation results associated with the models. Model variants include non-first-term observations vs. first-term observations, in-person observations vs. online observations, full set of predictors vs. admin-only and LMS-only predictors. We also test the models which only include certain subcategories of admin or LMS predictors.
* Figure 1, Figure 2, Table 2, Table 3, Figure 5 Appendix Table A2, Appendix Table A3, Appendix Table A6, Appendix Table A7 in the paper are generated by the scripts in this section.
* Run the scripts following the order outlined below.

**Group 1:**

1. **RF\_full.ipynb**:

* This script runs the random forest model using the full set of predictors, for all non-first-term courses. It also creates the predictor crosswalk as well as the data set that contains all predicted scores for all non-first-term observations.

1. **RF\_full\_firstterm.ipynb**:

* This script runs the random forest model using the full set of predictors, for all first-term courses. It also creates the data set that contains all predicted scores for all first-term observations.

1. **RF\_admin\_only.ipynb**:

* This script runs the random forest model only using the admin predictors, for all non-first-term courses.

1. **RF\_lms\_only.ipynb**:

* This script runs the random forest model only using the LMS predictors, for all non-first-term courses.

1. **RF\_admin\_only\_firstterm.ipynb**:

* This script runs the random forest model using only the admin predictors, for all first-term courses.

1. **RF\_lms\_only\_firstterm.ipynb**:

* This script runs the random forest model using only the LMS predictors, for all first-term courses.

1. **RF\_full\_num\_of\_predictors.ipynb**:

* This script is used to find out how model performance changes with number of predictors for the full-predictor, non-first-term model.

1. **RF\_full\_firstterm\_num\_of\_predictors.ipynb**:

* This script is used to find out how model performance changes with number of predictors for the full-predictor, first-term model.

1. **RF\_admin\_only\_num\_of\_predictors.ipynb**:

* This script is used to find out how model performance changes with number of predictors for the admin-only, non-first-term model.

1. **RF\_admin\_only\_firstterm\_num\_of\_predictors.ipynb**:

* This script is used to find out how model performance changes with number of predictors for the admin-only, first-term model.

1. **RF\_lms\_only\_num\_of\_predictors.ipynb**:

* This script is used to find out how model performance changes with number of predictors for the LMS-only, non-first-term model.

1. **RF\_lms\_only\_firstterm\_num\_of\_predictors.ipynb**:

* This script is used to find out how model performance changes with number of predictors for the LMS-only, first-term model.

**Group 2a:**

1. **RF\_full\_alternative.py**:

* This script fits the full-predictor RF model on all non-first-term observations without imputation of missing values -- just naturally treat missing values as one separate category.

1. **RF\_full\_firstterm\_alternative.py**:

* This script fits the admin\_only and lms\_only RF model on all non-first-term observations without imputation of missing values -- just naturally treat missing values as one separate category.

1. **RF\_admin\_lms\_only\_alternative.py**:

* This script fits the full-predictor RF model on all first-term observations without imputation of missing values -- just naturally treat missing values as one separate category.

1. **RF\_admin\_lms\_only\_firstterm\_alternative.py**:

* This script fits the admin\_only and lms\_only RF model on all first-term observations without imputation of missing values -- just naturally treat missing values as one separate category.

1. **RF\_full\_inperson\_vs\_online.ipynb**:

* This script fits the non-first-term model using exclusively in-person observations vs. the model using exclusively online observations.

1. **RF\_full\_firstterm\_inperson\_vs\_online.ipynb**:

* This script fits the first-term model using exclusively in-person observations vs. the model using exclusively online observations.

**Group 2b:**

1. **multinomial\_RF\_full.ipynb**:

* This script fits the full-predictor multinomial random forests model using all non-first-term observations.

1. **multinomial\_RF\_full\_firstterm.ipynb**:

* This script fits the full-predictor multinomial random forests model using all first-term observations.

**Group 3:**

1. **RF\_full\_inperson\_vs\_online.ipynb**:

* This script fits the non-first-term full-predictor model using exclusively in-person observations vs. the model using exclusively online observations.

1. **RF\_full\_firstterm\_inperson\_vs\_online.ipynb**:

* This script fits the first-term full-predictor model using exclusively in-person observations vs. the model using exclusively online observations.

**Group 4:**

1. **RF\_full\_upweighted.ipynb**:

* In this script, we fit the full-predictor RF model using all non-first-term observations through upweighting: multiplying each observation equal to the inverse of the frequency of its outcomes.

1. **RF\_admin\_only\_upweighted.ipynb**:

* In this script, we fit the admin-only RF model using all non-first-term observations through upweighting: multiplying each observation equal to the inverse of the frequency of its outcomes.

1. **RF\_lms\_only\_upweighted.ipynb**:

* In this script, we fit the lms-only RF model using all non-first-term observations through upweighting: multiplying each observation equal to the inverse of the frequency of its outcomes.

1. **RF\_full\_firstterm\_upweighted.ipynb**:

* In this script, we fit the full-predictor RF model using all first-term observations through upweighting: multiplying each observation equal to the inverse of the frequency of its outcomes.

1. **RF\_admin\_only\_firstterm\_upweighted.ipynb**:

* In this script, we fit the admin\_only RF model using all first-term observations through upweighting: multiplying each observation equal to the inverse of the frequency of its outcomes.

1. **RF\_lms\_only\_firstterm\_upweighted.ipynb**:

* In this script, we fit the lms\_only RF model using all first-term observations through upweighting: multiplying each observation equal to the inverse of the frequency of its outcomes.

**Group 5:**

1. **RF\_admin\_only\_subcategory1.ipynb**:

* Test the admin-only non-first-term model that only includes course-specific, subject-specific and instructor-related predictors.

1. **RF\_admin\_only\_subcategory2.ipynb**:

* Test the admin-only non-first-term model that only includes term-specific and demographic predictors.

1. **RF\_lms\_only\_subcategory1.ipynb**:

* Test the LMS-only non-first-term model that only includes the early-term predictors.

1. **RF\_lms\_only\_subcategory2.ipynb**:

* Test the LMS-only non-first-term model that only includes the prior-term predictors.

1. **RF\_lms\_only\_subcategory3.ipynb**:

* Test the LMS-only non-first-term model that only includes the prior early-term predictors.

1. **RF\_lms\_only\_subcategory4.ipynb**:

* Test the LMS-only non-first-term model that only includes the prior full-term predictors.

1. **RF\_lms\_only\_subcategory5.ipynb**:

* Test the LMS-only non-first-term model that only includes the early-term predictors that are related to the target course.

1. **RF\_lms\_only\_subcategory6.ipynb**:

* Test the LMS-only non-first-term model that only includes the early-term predictors that are related to concurrent courses.

1. **RF\_admin\_only\_firstterm\_subcategory1.ipynb**:

* Test the admin-only first-term model that only includes course-specific, subject-specific and instructor-related predictors.

1. **RF\_admin\_only\_firstterm\_subcategory2.ipynb**:

* Test the admin-only first-term model that only includes term-specific and demographic predictors.

1. **RF\_lms\_only\_firstterm\_subcategory1.ipynb**:

* Test the LMS-only first-term model that only includes the early-term predictors that are related to the target course.

1. **RF\_lms\_only\_firstterm\_subcategory2.ipynb**:

* Test the LMS-only first-term model that only includes the early-term predictors that are related to concurrent courses.

**03\_course-specific\_predictive\_models**

* There are five subdirectories under this directory: BIO 101, ENG 111, ENG 112, MTH 154, MTH 161. They are the five popular VCCS courses which we select to construct course-specific models
* Under each subdirectory (generically, suppose the course is SUB XXX), we include scripts that can be used to fit the random forests models using the study sample corresponding to the course SUB XXX, and generate the model performance evaluation results associated with the models. Model variants include full set of predictors vs. admin-only and LMS-only predictors. We also test the models which only include certain subcategories of admin or LMS predictors. Finally, we compare the performance of each course-specific model in terms of non-first-term observations vs. first-term observations, as well as in-person observations vs. online observations.
* Figure 3, Figure 4, Appendix Table A4, Appendix Figure A1 in the paper are generated by the scripts in this section.
* Run the scripts under each subcategory (generically, suppose the course is SUB XXX) following the order outlined below.

1. **RF\_SUBXXX\_full.ipynb**:

* This script fits the random forest model for the course SUB XXX, using the full set of predictors.

1. **RF\_SUBXXX\_admin\_only.ipynb**:

* This script fits the random forest model for the course SUB XXX, using only the admin predictors.

1. **RF\_SUBXXX\_lms\_only.ipynb**:

* This script fits the random forest model for the course SUB XXX, using only the LMS predictors.

1. **RF\_SUBXXX\_admin\_only\_subcategory1.ipynb**:

* This script tests the SUBXXX-specific model which only includes the course-specific & instructor-related & subject-specific admin predictors.

1. **RF\_SUBXXX\_admin\_only\_subcategory2.ipynb**:

* This script tests the SUBXXX-specific model which only includes the term-specific & demographic admin predictors.

1. **RF\_SUBXXX\_lms\_only\_subcategory1.ipynb**:

* This script tests the SUBXXX-specific model which only includes the early-term LMS predictors in the target term.

1. **RF\_SUBXXX\_lms\_only\_subcategory2.ipynb**:

* This script tests the SUBXXX-specific model which only includes the LMS predictors in prior terms (including both early-term and full-term ones).

1. **RF\_SUBXXX\_lms\_only\_subcategory3.ipynb**:

* This script tests the SUBXXX-specific model which only includes the early-term LMS predictors in prior terms.

1. **RF\_SUBXXX\_lms\_only\_subcategory4.ipynb**:

* This script tests the SUBXXX-specific model which only includes the full-term LMS predictors in prior terms.

1. **RF\_SUBXXX\_lms\_only\_subcategory5.ipynb**:

* This script tests the SUBXXX-specific model which only includes the early-term LMS predictors associated with the target course in the target term.

1. **RF\_SUBXXX\_lms\_only\_subcategory6.ipynb**:

* This script tests the SUBXXX-specific model which only includes the early-term LMS predictors associated with the concurrent courses in the target term.