Data and Programs for: *Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education.*

Kelli A. Bird, Benjamin L. Castleman, Zachary Mabel & Yifeng Song

July 16th, 2021

This memo documents the analysis files used to generate the tables/figures found in “Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education” in AERAOpen, 2021.

**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)

Due to the nature of the student-level VCCS data, we will not be able to share the data files and upload them to the OpenICPSR repository.

**Program Code Files for the Analysis**

By execute the scripts in the 19 subdirectories listed below in order, all of the tables/figures in this paper can be generated:

**(1) build\_data**

Run the script “**00\_master\_script.do**” to perform necessary initial cleaning steps for the raw VCCS administrative data files and prepare external data to be merged into the VCCS data files. This .do file calls all other scripts under this subdirectory and executes them in order.

**(2) master\_student\_x\_term\_data**

Run the script “**00\_master\_build\_student\_x\_term.do**”, which calls all other scripts under this subdirectory to reorganize, aggregate and manipulate the build VCCS files created in the previous step, and then generate the master student x term observations that contain key information necessary for creating the study sample of this paper.

**(3) study\_sample\_construction**

* First run the script “**01\_sample\_restrictions.do**” to identify the list of students to be included in the study sample based on the selection criteria described in II.C and Appendix 1 of the paper.
* Next run the script “**02\_constructing\_sample\_and\_outcome.do**” to create the outcome of interest for each student in the study sample, based on the II.B of the paper.

**(4) preprocessing**

Run the **Script 01-07** in order under this subdirectory following their indices. Those scripts are key preprocessing steps for cleaning/aggregating the VCCS data that have to do with course-taking records and GPA, and for generating the demographic and term-level academic predictors (e.g. term GPA, cumulative GPA, trend of term-level enrollment intensity).

**(5) non\_truncated\_predictors**

Run the script “**construct\_full\_dataset.do**” to construct all of the predictors for the non-truncated sample, and then complete the training/validation splitting of the dataset, as described in II.B and II.C of the paper.

**(6) truncated\_predictors**

* First execute the **Script 01-05** in order under this subdirectory to complete the random truncation procedure for the study sample: for each student in the study sample, we assign an end term for the post-truncation observation window in such a way that the both the truncated training and validation sample closely resemble the currently enrolled cohort. Those scripts are relevant to the creation of Table 1, Appendix Table A4.
* Next run the script “**06\_construct\_full\_dataset\_truncated.do**” (This script is similar to “construct\_full\_dataset.do” under subdirectory 5) to construct all of the predictors for the randomly truncated sample, which will be used to fit the six base models except for RNN. Then run the script “**07\_construct\_full\_dataset\_coxph.do**” to slightly modify the dataset created by the previous script such that the data format could fit into the CoxPH modeling.

**(7) rnn\_data\_construction**

* First run the script “**01\_clean\_term\_specific\_part.do**” to reshape the term-specific part of the truncated dataset created by the previous subdirectory into a format such that each term-specific predictor has a timestamp that can be fed into an RNN model.
* Then run the **Script 02-04** in order under this subdirectory to construct the input data (multi-dimensional arrays) for fitting the RNN model for the base model, the RNN model without demographic predictors and the RNN model that excludes the complexly specified term-specific predictors.

**(8) base\_models\_fitting**

* First run the **Script 01-04** in order under this subdirectory to fit the five base models (OLS, Logit, CoxPH, RF and XGBoost) on the training data, obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Figure 2, 3, 9 and Appendix Table A1 of the paper.
* Next run the **Script 05-06** in order under this subdirectory to perform automated feature selection (using lasso logistic regression) as well as generate the feature importance ranking from the OLS and Logit base models. The output of those scripts are relevant to Appendix Figure A18 of the paper.
* Then run the script “**07\_Overlapping\_in\_feature\_ranking.ipynb**” to identify the commonality of top 20% of important features across base models and generate Figure 8 of the paper.
* Finally run the **Script 08-09** in order under this subdirectory to clean up the output from the previous scripts under this subdirectory, and generate Appendix Table A5 and A6 of the paper.

**(9) rnn\_modeling**

* The **Script 01a-01g** is the recorded procedure of fine-tuning the base RNN model in order to identify the optimal architecture and hyperparameters (see Appendix 3.(3) for more details). In practice, when applied to a different training sample, it does not have to follow the order specified by this set of scripts.
* Next run the script “**02\_run\_lstm\_model\_no\_demo.py**” to fit the RNN model for the training sample that do not include demographic predictors. For simplicity purposes we apply the optimal RNN model setup we identified from the base model. This model will be used to compare against the model variants of subdirectory 11.
* Then run the script “**03\_run\_lstm\_model\_simple.py**” to fit the RNN model for the training sample that exclude the complexly specified term-specific predictors. For simplicity purposes we apply the optimal RNN model setup we identified from the base model. This model will be used to compare against the model variants of subdirectory 12.
* Finally run the **Script 04-06** in order under this subdirectory to obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Figure 2, 3, 10; Appendix Table A1; Appendix Figure A8 of the paper.

**(10) base\_models\_evaluation**

* First run the script “**01\_Comparing\_Predicted\_Scores.ipynb**” to perform comparisons of relative risk rankings (predicted scores) across the six base models (section III.D of the paper). Specifically, this script generates the Figure 5, 6.1, 6.2, 6.3, 7; Appendix Table A2; Appendix Figure A1-A7 of the paper.
* Then run the script “**02\_Comparing\_Predicted\_Scores\_2.ipynb**” to perform comparison of the consistency of predicted outcome (binary) across the six base models and generate Figure 4 of the paper.

**(11) truncated\_without\_demo\_predictors**

* First run the **Script 01-04** in order under this subdirectory to fit the five models without demographic predictors (OLS, Logit, CoxPH, RF and XGBoost) on the training data, obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Table 2 and Figure 10 of the paper.
* Next run the script “**05\_RFE.ipynb**” to generate the feature importance ranking from the OLS and Logit models that exclude demographic predictors. The output of this script will be used in the next script.
* Then run the script “**06\_Comparing\_Predicted\_Scores\_New.ipynb**” to test the effect of excluding demographic predictors on the feature importance ranking and the stability of relative risk rankings. Specifically, this script generates Appendix Table A3 and Appendix Figure A21, A22.

**(12) truncated\_simple\_predictors**

* First run the **Script 01-04** in order under this subdirectory to fit the five models without complexly specified term-specific predictors (OLS, Logit, CoxPH, RF and XGBoost) on the training data, obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Figure 2; Appendix Table A1 and Appendix Figure A8 of the paper.
* Then run the script “**05\_Comparing\_Predicted\_Scores\_New.ipynb**” to test the effect of excluding complexly specified predictors on the stability of relative risk rankings. Specifically, this script generates Appendix Figure A9 and A10.

**(13) truncated\_simple\_predictors\_2**

* First run the **Script 01a-04b** in order under this subdirectory to fit the five models that only include simple non-term-specific predictors (OLS, Logit, CoxPH, RF and XGBoost) on both the non-truncated and the truncated training data, obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Figure 1, 2; Appendix Table A1 and Appendix Figure A14 of the paper.
* Then run the script “**05\_Comparing\_Predicted\_Scores\_New.ipynb**” to test the effect of excluding complexly specified non-term-specific predictors and all term-specific predictors on the stability of relative risk rankings. Specifically, this script generates Appendix Figure A15 and A16.

**(14) truncated\_without\_term\_specific**

* First run the **Script 01a-04b** in order under this subdirectory to fit the five models that only include all of the non-term-specific predictors (OLS, Logit, CoxPH, RF and XGBoost) on both the non-truncated and the truncated training data, obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Figure 1, 2; Appendix Table A1 and Appendix Figure A11 of the paper.
* Then run the script “**05\_Comparing\_Predicted\_Scores\_New.ipynb**” to test the effect of excluding all term-specific predictors on the stability of relative risk rankings. Specifically, this script generates Appendix Figure A12 and A13.

**(15) truncated\_reduced**

* First run the **Script 01-04** in order under this subdirectory to fit the five models which only include the 147 predictors identified by the lasso feature selection method (OLS, Logit, CoxPH, RF and XGBoost) on the training data, obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Appendix Figure A17 of the paper.
* Then run the script “**05\_Comparing\_Predicted\_Scores\_New.ipynb**” to test the effect of excluding predictors dropped by the feature selection method on the stability of relative risk rankings. Specifically, this script generates Appendix Figure A19 and A20.

**(16) smaller\_training\_sample**

* Subfolder “**test1**”:
  + First run the **Script 01-04** in order under this subdirectory to fit the five models without complexly specified term-specific predictors (OLS, Logit, CoxPH, RF and XGBoost) on the reduced training data (which consists of a random sample of 10% of the original training data), obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Appendix Figure A24 of the paper.
  + Then run the script “**05\_Comparing\_Predicted\_Scores\_New.ipynb**” to test the effect of reducing the training sample size by 90% on the stability of relative risk rankings. Specifically, this script generates Appendix Figure A27 and A28.
* Subfolder “**test2**”:
  + First run the script “**01\_create\_pvcc\_ind.do**” to identify all students in the study sample who enrolled in PVCC during the last term in the truncated time window. The rest of the analyses in this subsection will be focused on this subgroup of students.
  + Next run the **Script 02-04** in order under this subdirectory to fit the four models without complexly specified term-specific predictors (OLS, Logit, RF and XGBoost) on the PVCC-only training data, obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Appendix Figure A23 of the paper.
  + Then run the script “**05\_Comparing\_Predicted\_Scores\_New.ipynb**” to test the effect of restricting the training sample to PVCC students vs. using the VCCS system-wide training sample on the stability of relative risk rankings. Specifically, this script generates Appendix Figure A25 and A26.

**(17) truncated\_without\_nsc\_predictors**

* First run the **Script 01-04** in order under this subdirectory to fit the five models which excludes the NSC predictors (OLS, Logit, CoxPH, RF and XGBoost) from the training data, obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Appendix Figure A32 of the paper.
* Then run the script “**05\_Comparing\_Predicted\_Scores\_New.ipynb**” to test the effect of excluding NSC predictors on the stability of relative risk rankings. Specifically, this script generates Appendix Figure A33 and A35.

**(18) truncated\_without\_nsc\_enrollees**

* First run the **Script 01-04** in order under this subdirectory to fit the five models which excludes the all of the ever NSC enrollees (OLS, Logit, CoxPH, RF and XGBoost) from the training data, obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Appendix Figure A32 of the paper.
* Then run the script “**05\_Comparing\_Predicted\_Scores\_New.ipynb**” to test the effect of excluding NSC predictors on the stability of relative risk rankings. Specifically, this script generates Appendix Figure A34 and A36.

**(19) truncated\_alternative\_definitions**

* First execute the **Script 01-03** in order under this subdirectory to complete the random truncation procedure for the study sample using the alternative: for each student in the study sample, we assign an end term for the post-truncation observation window in such a way that the both the truncated training and validation sample closely resemble the currently enrolled cohort. Since the alternative outcome definition excludes all NSC data, the graduation time might change according to our definition. Therefore, it is necessary to re-run the procedure of random truncation as we did for the base models.
* Next run the script “**04\_construct\_full\_dataset\_truncated.do**” (This script is similar to Script 06 under subdirectory 6) to construct all of the predictors for the randomly truncated sample using the alternative outcome definition, which will be used to fit the five models (OLS, Logit, CoxPH, RF and XGBoost). Then run the script “**05\_construct\_full\_dataset\_coxph.do**” to slightly modify the dataset created by the previous script such that the data format could fit into the CoxPH modeling (This script is similar to Script 07 under subdirectory 6).
* Then run the **Script 06-09** in order under this subdirectory to fit the five models (OLS, Logit, CoxPH, RF and XGBoost) on the training data using the alternative outcome definition, obtain the key evaluation metrics (c-statistics, F1-score, etc.) on the validation sample, and generate the predicted scores for each observation in the validation sample which will be used for subsequent analyses. The output of those scripts are relevant to Appendix Figure A29 of the paper.
* Then run the script “**10\_Comparing\_Predicted\_Scores\_New.ipynb**” to test the effect of using the alternative outcome definition on the stability of relative risk rankings. Specifically, this script generates Appendix Figure A30 and A31.
* Finally run the script “11\_Comparing\_Model\_Performance.ipynb” to