Machine Learning Modeling to Predict At-Risk Students

Jim Creighton, Albert Lee, Max Tsisyk

Introduction

We used the University of Michigan Learning Analytics Data Architecture (LARC) data set, containing both demographic and academic metrics, to predict academically "at-risk" undergraduate students. We approached this through developing multiple models and creating multiple feature sets to identify those combinations which created highest predictive performance.

Data Set

The LARC data set is a research-focused data set capturing student-level data at University of Michigan - Ann Arbor (anonymized data with unique student identifiers). It encompasses records for more than 390,000 students, with more than 600 recorded data points (features) per student. From the May 25, 2021 update, we chose to focus on:

- **Student Data**: Data which is *constant* throughout a student's academic career (e.g. ethnicity, SAT test scores, high school GPA, family income, etc.). This data is independent of the student's college experiences. This data can commonly be thought of as the information to which an admissions officer might possess when reviewing a student admissions application. (441 Features; 390,000 records)
- **Student-Term Data**: Data which can *change from term to term* (e.g. academic career choice such as LSA vs. Law vs. Engineering, majors and minors, term GPA, athletic team participation, etc.). This data is dependent upon the student's college experiences. (160 Features; 2.8M records)

Additionally, we acquired external census data and used the student zip codes to map demographic information pertaining to the student's address. For the zip code associated with each student, we were able to input the amount of people covered by insurance, percentage on food assistance, median household income, access to transportation, family size, etc. While this did not give us information directly related to the individual student, it did give us some additional information about the potential environment that the student came from.

Scope

The original LARC data had student records back to 1982. We chose to narrow to (A) undergraduate students, at (B) UM Ann Arbor, (C) who matriculated any time in 2012-15. This reduced the initial data set to approximately 30,000 unique **Student** and 250,000 **Student-Term** (semester) records. The choice of narrowing to 2012-15 matriculation was a balance between a desire for data recency and the fact that our predominant "at-risk" definition required at least five years of post-matriculation data (see below).

Defining the "At-Risk" Prediction Label

Though we initially considered numerous approaches to defining a predictive label to identify an academically at-risk student, data limitations caused us to relatively quickly narrow to two main approaches:

1. A student permanently withdrawing and/or failing to graduate within 5 years
2. A student's *term* GPA ever dropping below 2.0 (a "C" grade)

Several subsequent back-and-forths with the data owners (the university registrar office) helped us better understand how specific data fields were defined and recorded. Importantly, this uncovered that:

- While withdrawal from an individual term is captured, LARC data doesn't record permanent withdrawal from the university
- Term GPAs can be low (including 0.0) for artificial reasons that don't indicate academic problems

With the registrar's feedback, we were able to calculate/create at-risk labels using both intended approaches, though we dropped utilizing the GPA-based label about halfway through our work (A) due to recording inconsistencies described above, (B) because its use prevented using any other GPA-related features in the model, and (C) to reduce the total number of evaluations required. In the end, our main at-risk label (to predict) was "Failure to graduate within 5 years", which could be cleanly calculated from the LARC data.

Modelling Approach

We built supervised machine learning models to classify (predict) each student as academically "at-risk" or "not at-risk". The initial goal was to compare multiple supervised classification model types (Logistic Regression, Gaussian Naive Bayes, Random Forest Classifier, Multi-layer Perceptron, Decision Tree Classifier) and compare three different feature sets. Based upon performance metrics, we narrowed to one model type and then further developed the models. The three feature sets compared were selected from the:

1. Independent features (i.e. "Independent of the college experience"; from the **Student Data**)
2. Dependent features (i.e. "Dependent upon the college experience"; from the **Student-Term** Data)
3. Independent & Dependent features (combination of both above)

Once these comparisons were complete, the "winning" models were further explored for fairness.

Workflow

Our workflow encompassed numerous Jupyter notebooks (translated to a DVC pipeline) to preprocess data, model data, and evaluate the models. Data was passed to subsequent notebooks with standardized CSV formats. A GitHub repository was set up to manage changes during development.

GitHub Repository

A DVC pipeline was established from the notebooks to ensure up-to-date data was being utilized, allow testing of multiple variations of the features and models, and enable reproducibility. By utilizing a DVC pipeline, we ensured that all computational steps used to produce the model from the data were captured and easily reproduced. See Appendix A for the DVC pipeline / workflow.

Pre-Processing

The Student Data (data independent of the college experience; one record per student) and Student-Term Data (data dependent upon the college experience; one record *for each term* the student attended) are maintained by LARC in separate flat files and have their own unique features. Therefore, each set of data must be processed separately.

Feature Selection

To make the models more manageable, especially since we knew we'd need to one-hot-encode the many categorical features), we did an initial cut of the 600+ features. Student Data and Student-Term Data features were reduced, separately, by taking similar approaches:

- Removing features that would contain almost identical / redundant information
- Identifying highly sparse features, and removing those when we believed had little information value (see example results below). See Appendix B for a sparse variable overview.
- Removing features that, in the context of our goal, we believed had minimal information value

This process reduced the Student Data from 440 to 61 features and the Student-Term Data from 160 to 16. Categorical variables were then one-hot-encoded, which expanded the feature set again to 335 and 41, respectively. During model training, we applied feature pruning to remove features that did not improve model performance, and this reduced the feature set to 248 and 36 features.

	All Original Features	Initial Feature Selection	Post Feature Encoding	Post Feature Pruning*
Student Data	440	61	335	248
Student Term Data	160	16	41	36

**Displaying features used for the best performing classifier (Random Forest)*

Train/Validation/Test Split Approach

The notebooks and pipeline were set up with a 80 / 10 / 10 train, validation, test split on the initial LARC **Student** and **Student-Term** data. Importantly, this split was conducted *before* we did any filling of missing values with means (or other appropriate functions) so that we'd avoid data leakage.

Missing Values

The remaining selected features had many missing values which needed to be filled, since we didn't want to drop these records prior to modelling. We had to investigate the meaning of missing values in each field, as it varied by field. We selected a method to fill each field (like mean, max, zeroes, etc.) based upon what logically made sense in the context of the model and field type (categorical/continuous). Additionally, all filling of missing values was conducted after the train/test split to avoid data leakage. As an example, the recorded TOEFL score (an English Fluency test) is a null value if the test wasn't taken (wasn't required), which would be the case for native English speakers. Since these native speakers would likely be fluent, we therefore filled this value with the max value of the TOEFL scores achieved.

Student-Term Data Feature Engineering

One of the overall goals was to predict whether a student was academically at-risk *based upon their term data*. Since students have a separate record (containing a unique measurement of each feature) for *each term*, we created a method to collapse all of those *term* records into a single "aggregated" record for each student. Specifically, we created that single record by taking aggregations like the max, min, mean, standard deviation, and/or unique count (for categoricals), etc. for each feature across all of a student's *term* records. In this fashion, we could create a single record for the student, reflecting the data from all of their terms, and could use the same pipeline into our supervised models.

Model Exploration

Each of the three feature sets were initially utilized with each of five supervised classification model types (Logistic Regression, Gaussian Naive Bayes, Random Forest Classifier, Multi-layer Perceptron (MLP), Decision Tree Classifier). We used the corresponding classes from the scikit-learn library (Logistic Regression, GaussianNB, RandomForestClassifier, and MLPClassifier). An initial side-by-side comparison of the five models (with same combined Independent & Dependent features) allowed us to discard the Multi-layer Perceptron model as it performed the most poorly out of all the models. After an initial sanity check, we discarded the Decision Tree Classifier since it was of the same family as the Random Forest but consistently returned lower-scoring predictions. Initial AUC scores were:

- Logistic Regression: 0.606
- Gaussian Naive Bayes: 0.671
- Random Forest Classifier: 0.768
- MLP Classifier: 0.568

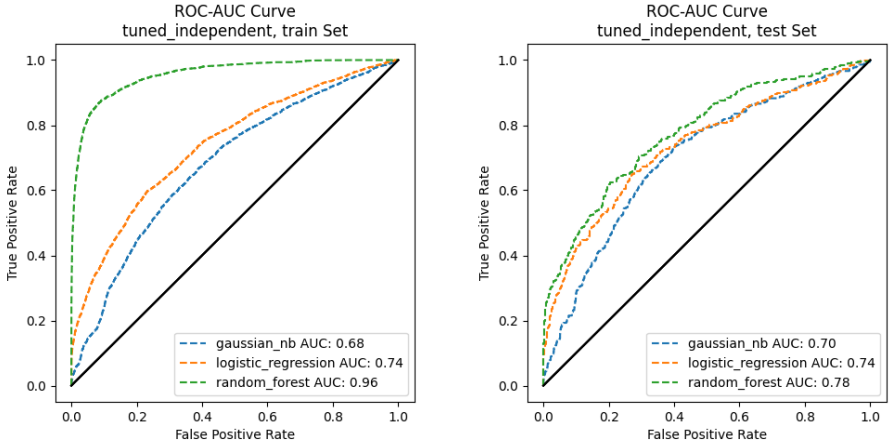
Therefore, knowing that feature pruning and hyperparameter tuning for each model/feature set combination (especially for the MLP) would take significant time, we cut out the MLP Classifier and Decision Tree Classifier from further consideration.

With the remaining three models and three feature sets (nine combinations total), we then performed feature pruning (on an automated basis) by evaluating individual feature importance on each model / feature set. To do this, we repeatedly calculated the permutation importance of all features using the validation data and dropped the least important feature. We continued this process until all remaining features showed positive importance.

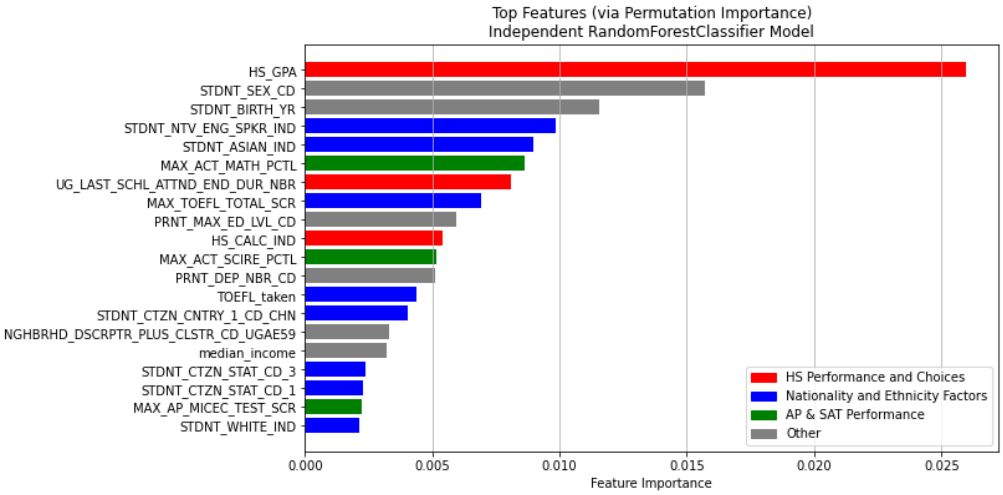
Further, once features were pruned, we then performed parameter grid search (5-fold cross validation) for each model / feature set. The incremental model performance (AUC) improvements through feature pruning and hyperparameter tuning are available in Appendix C. We were surprised to find that the performance improvements from pruning and tuning were marginal (did not significantly improve AUC over the baseline models).

A) Independent Features (i.e. features “Independent of the college experience”; from the Student Data)

ROC AUC performance was used as the basis for the Independent feature model evaluation/comparison and is shown below.

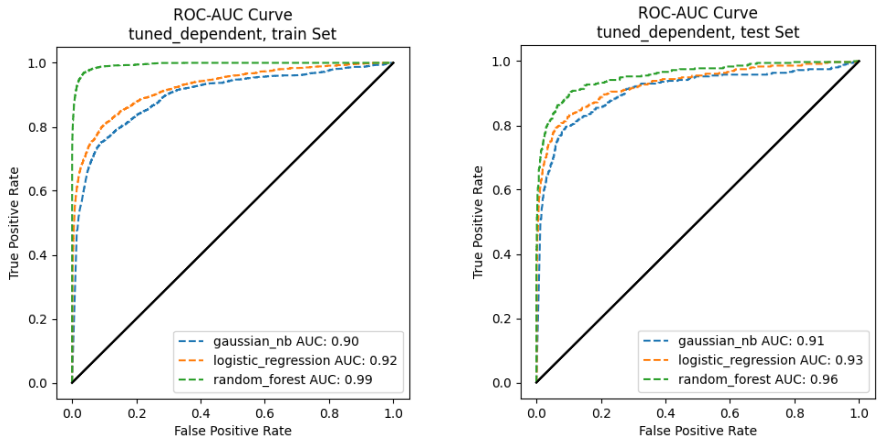


It’s clear that the Random Forest Classifier outperforms the models on the same set of features, though it is only marginally better than the Logistic Regression. Through grid search, we found the optimal Random Forest model had 100 trees with no limit on depth. From the leading Random Forest Classifier model, we evaluated feature importance (in Appendix D) to get a sense of which features contain the most information value.

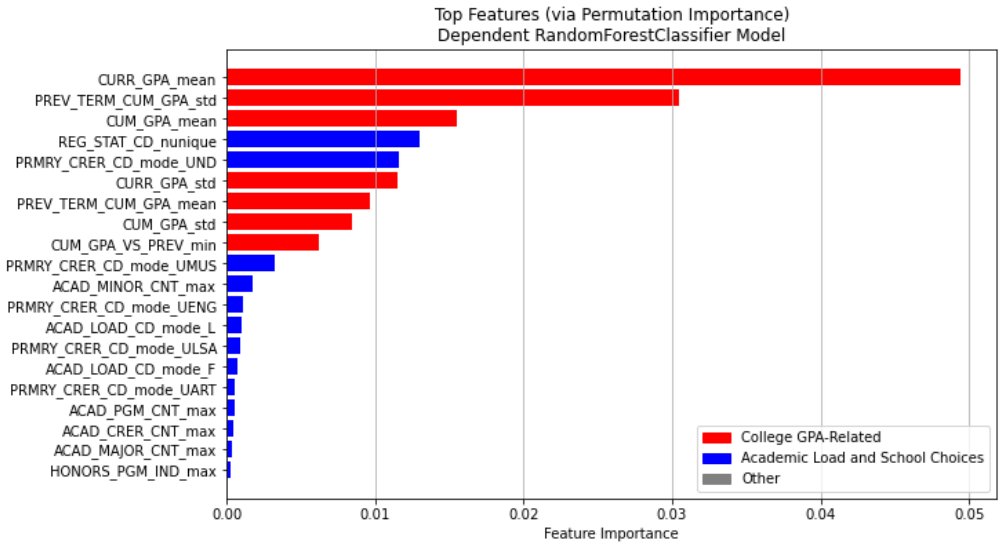


B) Dependent Features (i.e. features “Dependent upon the college experience”; from the Student-Term Data)

As described earlier, the features from all of the student’s terms were collapsed (aggregated with a metric such as mean) into a single record for each student. ROC AUC performance was used as the basis for the Dependent feature model evaluation/comparison and is shown below.

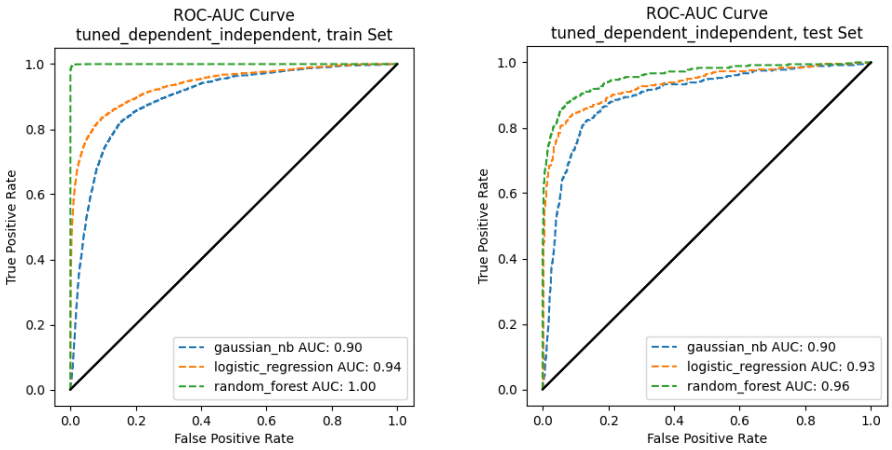


The Random Forest Classifier again outperforms the other models on the same set of features, with a slightly larger lead over the Logistic Regression. Through grid search, we found the optimal Random Forest model had a depth of 25 with 125 trees. We again evaluated feature importance of the Random Forest Classifier to get a sense of which features contain the most information value.

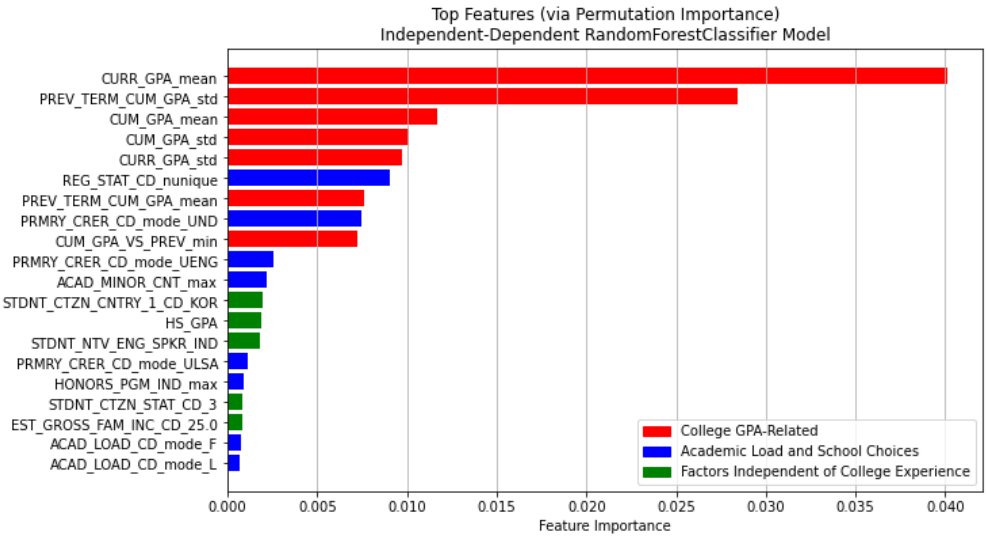


C) Independent & Dependent Features (combination of both above)

While it was valuable to understand model performance when developed utilizing solely the Dependent features, it is likely that anyone (counselor, administrator, etc.) with access to the Dependent feature information about a student will also have access to the Independent features as well. We therefore wanted to evaluate models utilizing both the dependent and Independent features. The ROC AUC performance for such models is shown below.



Yet again, the Random Forest Classifier outperforms the other models on the same set of features. However, the addition of the Independent features to the Dependent features creates only a very slight improvement over the Dependent-only model. The increase is so small that it only appears in the third decimal place. Therefore, from a practical standpoint, it *may not always be worth* the additional complexity of incorporating the Independent features into the model for such a small improvement (however, we do use the Independent/Dependent model in further analysis since it does, in fact, perform better). Through grid search, we found the optimal Random Forest model had a depth of 25 with 125 trees Looking across the combined feature sets, we evaluated feature importance.

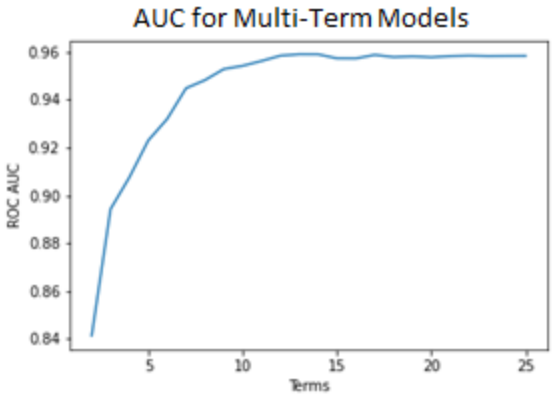


An Alternate Independent/Dependent Features Models

Models based upon the Independent feature sets provide an one-time opportunity to predict whether a student will be academically at-risk, based upon this historical data which is independent of their actual college experiences. Such a prediction could be valuable as a method for the admissions office to flag a student as one "to watch" more closely and/or to proactively provide additional academic assistance.

Our initial Independent/Dependent feature model, discussed immediately above, collapsed a student's total college career (all terms) into one record. While this model performed well, it is based upon having *all* of a student's term data available, which means we'd be predicting whether they are at-risk (whether they would graduate in 5 years) only after-the-fact. While this is an intellectually interesting question, its real-world value is limited, since it's then too late to intervene. There's more real-world value to be able to predict, for example, whether a student that has finished three terms is at-risk of not graduating. This would provide an opportunity for faculty to intervene and provide academic support. We therefore also took a second, slightly different, approach with the Dependent feature model(s). Specifically, we trained separate models for each accumulated number of terms. In other words, we trained one model on only the first two terms' worth of data (using aggregations such as mean or max across the two terms), then another model using the first three terms, then another for the first four terms, and so on. The implication is that a student who just completed his third term could be flagged as at-risk of not graduating (in 5 years) by using the "first-three-term" model.

The ROC AUC performance of each term model, from 2 to 25 terms (25 is the maximum of any student) is shown below. Note: Because of the consistent better performance of the Random Forest Classifier in the prior analyses, only a Random Forest Classifier was evaluated here. Additionally, to avoid the time-consuming feature pruning and hyperparameter tuning for each of the "term" models, we utilized the previously-found best hyperparameters for the Independent/Dependent Random Forest Classifier. It's notable that even at just 2-3 terms into their career, the models achieve a relatively respectable predictive ability (AUC score in the 0.84-0.90 range). However, the predictive ability rises even higher (up to ~0.96) in later terms.



Implications of Identified Feature Importance

In the models discussed above, several features were frequently identified as having high importance. High importance means that they influence the model's prediction. In addition, we view high importance variables as important signals which administrators may want to monitor as signs of a student being at-risk. (However, the models do not necessarily represent the real world exactly, so high importance does not definitively mean that the feature, individually, is a sign of risk. These features may be worth studying more to establish a definitive link). We additionally performed Shapley value (see Appendix D) and partial dependence analysis to understand whether the features were positively (or negatively) associated with risk of not graduating within 5 years. Important features include:

Independent Features:

- High School Performance - A student's high school GPA (HS_GPA) was found to be the leading Independent predictive feature, by far. Though this may not be surprising to an experienced educator, it suggests that students with low high school GPAs might be watched and/or provided with proactive support early-on in their college career. Though less predictive, knowing that a student has taken calculus in high school (HS_CALC_IND) is important as well. Higher GPA and taking Calculus are negatively associated with the risk of not graduating.
- Family Characteristics - Family income (EST_GROSS_FAM_INC_CD) and parent educational achievement level (PRNT_MAX_ED_LVL_CD) were both top-10 features in predicting 5-year graduation, and are both negatively associated with risk of not graduating. The percent of students receiving public assistance at the student's school (public_ass_perc, a census-added feature) was only slightly less important and is positively associated with the risk of not graduating. These all argue for ensuring this data is collected for all admitted students. Further, though it is not definitive (and it should be studied more to establish an actual link), it suggests that students with lower family incomes or educational levels should be watched more closely once enrolled.
- Foreign Student Attributes - Whether a student had taken the TOEFL test (a binary indicator derived from the reported scores) was found to be a top-10 feature in predicting risk and is positively associated with risk. Citizenship (STDNT_CTZN_CNTRY_1_CD) or permanent residency (FRST_FRGN_PRMNNT_RES_CNTRY_CD) from specific countries, such as China and Korea (both positively associated with risk), were found to be important, though not in the top 10.
- Ethnicity - Having Asian ethnicity (STDNT_ASIAN_IND) was found to be a top-10 predictor of graduation risk (negatively associated with risk), regardless of actual citizenship or permanent residency.

Dependent Features:

- College GPA-Related Features - Features such as current term GPA (CURR_GPA_mean), cumulative GPA (CUM_GPA_mean) and GPA variability (PREV_TERM_CUM_GPA_std and CURR_GPA_std) are consistently ranked top-10 in importance. Since GPA is, by definition, an assessment of academic achievement, it intuitively makes sense that these metrics would contain high information value. Despite an "obvious" association between high GPA and likelihood to graduate, we felt it was important to leave these in as features (A) to emphasize the importance of administrators watching student GPAs as an indicator and (B) because there may well be cases

where students with high GPAs don't graduate. Notably, when GPA-related metrics were removed from the Independent/Dependent Feature Random Forest Classifier, AUC (test) dropped from 0.958 to 0.877. Further, we constructed a new metric for the size of the change in GPA from one term to another as a feature (CUM_GPA_VS_PREV). This change metric additionally showed in the top 10 features, but not as important as the other LARC-captured GPA metrics. Unsurprisingly, the actual GPA *values* were negatively associated with risk (generally, especially when below approx 3.4) and higher GPA *variability* was (generally) positively associated with risk. The implication, overall, is that GPA metrics should be tracked as signs of risk.

- Registration Status (REG_STAT_CD_nunique) - LARC data tracks whether a student was "Registered", "Cancelled" or "Withdrawn" in an *individual term*, but it doesn't track whether a student stopped taking classes *completely* (officially withdrew) from the university. This field (which counts a student's unique status codes) is identified as another high-importance feature, one which is positively associated with risk. Intuitively it makes sense that a student who doesn't graduate (in 5 years) might have had to withdraw/cancel a term at some point. Still, we chose to leave this feature in the set since cancellation/withdrawal from an individual term doesn't always mean that a student won't graduate (i.e., it's not data leakage). It just as easily could be a case that a student registered for summer term and then cancelled because a great internship was offered. Nevertheless, the models tell us that Registration Status might be a signal of potential problems and (after further study to establish a definitive link) it may be valuable for administrators to watch.
- Undecided and Declared Students (PRMRY_CRER_CD_mode_UND) - The fact that a student has failed to declare an academic career ("Undecided") is a top-10 predictor of graduation risk, and is positively associated with risk. This suggests that early intervention or coaching may aid these students. Notably, the specific school a student declared (Music, Engineering, LSA, Art) also had predictive ability, but were beyond the top-10.
- Academic Load (ACAD_LOAD_CD_mode_L and _mode_F) - LARC data tracks a few different categories indicating the degree to which a student is attending with a full time, or less, class load in a given term. These include "Full-time" (_mode_F), "Less-than-½-time" (_mode_L) and other intermediate designations. One-hot encoded versions of this feature, specifically Full-time and Less-than-½-time, show some (not top-10) importance in the prediction of whether a student is at-risk. Unsurprisingly, being mostly "Full-time" (_mode_F) is negatively associated with risk, while "Less-than-½-time" (_mode_L) is positively associated with risk.
  Because a student who chooses to attend "part-time" (anything lesser than full-time) might inherently take longer to graduate (which is related to our at-risk definition), we considered dropping this feature from the model. However, upon further investigation, we discovered that nearly 60% of all undergraduates become less-than-fulltime at some point in their career and, of those, only 12% don't graduate in 5 years (versus 10% of all students). In fact, manual examination indicated that it's common for students to become less-than-fulltime just in their final term(s). Therefore, we concluded that, in this context, becoming something other than full-time does not inherently force someone to take longer to graduate.

Continued Exploration of Leading Model: Model Fairness and Bias

Selection of Leading Models for Continued Exploration

As previously discussed, the Random Forest Classifier model outperformed other classifiers and therefore became our focus for further exploration. We felt it was important to continue pursuing *both* one model variation that consisted of Independent Features (independent of the college experience) and one including Dependent Features (dependent upon the college experience).

- Independent Features Model - We only had one such Random Forest Classifier (test set AUC score: 0.78), and continued with this.
- Independent/Dependent Features Model - As mentioned previously, anyone with access to the Dependent Features will likely also have access to the Independent Features. Since the combined model marginally outperformed the Dependent-only model (both approximately AUC score: 0.96), we continued with the combined Independent/Dependent Feature model instead. Note that, while the "alternate" Ind/Dep Feature model (discussed above) took a term-by-term approach to modelling and prediction, it also created two dozen models which we'd need to further evaluate. We therefore chose to continue our development with the Ind/Dep model containing all term records, which we viewed as a proxy for the many individual "term-level" models.

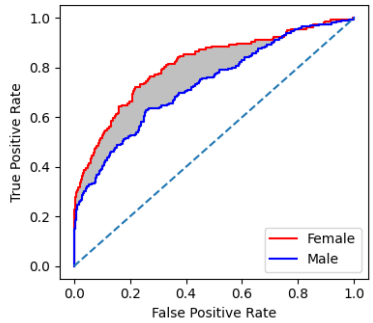
Model Fairness Evaluation - ABROCA Plots

While a model can have overall good performance, it can also perform meaningfully differently between subpopulations. One way to assess model fairness (among subpopulations) is to calculate the Absolute Between-ROC Area (ABROCA). This metric is an indication of the difference in model performance between groups, with a larger value indicating a greater difference in performance. The metric is the shaded area between the ROC curves of the model for each group. We plotted ABROCA for our best-performing Random Forest models.

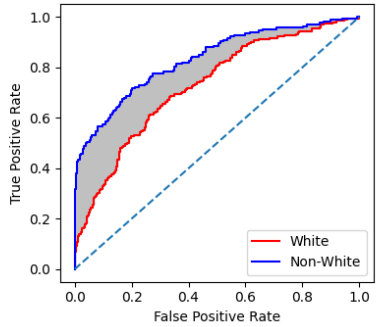
While there are many subgroups within the student data that could be tested, several were of particular interest to us -- Gender (M/F), Ethnicity (White/Non-white), and Residency (US/International).

- Independent Features Model - The ABROCA between Female/Male (0.08) was minimal in predicting which students would not graduate in 5 years. The models performed better for Non-White students than for White students and resulted in a higher ABROCA of approximately 0.10. The models performed better for International students than Non-International ones, resulting in an ABROCA of 0.19 on the test set (0.18 on the validation set). The better performance for Non-White and International students surprised us since these subpopulations had fewer total training points than their White and Non-International counterparts. However, upon investigation, we found that the Non-White and International subpopulations had better balance (between those that failed to graduate and those that graduated) than their counterparts, which may have led to better training of models. The end result of the ABROCA evaluation is that we believe that the Independent Features model is relatively fair across genders and White/Non-White, but less fair (performs more poorly) for Non-International students.

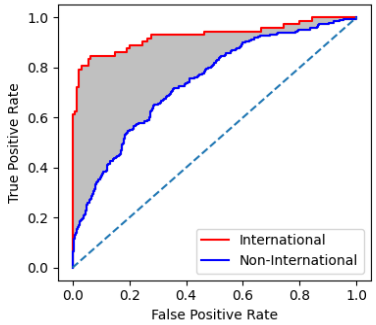
Gender Split
ABROCA: 0.08061



Race Split
ABROCA: 0.10424

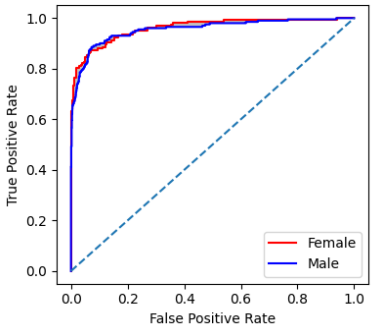


Nationality Split
ABROCA: 0.18570

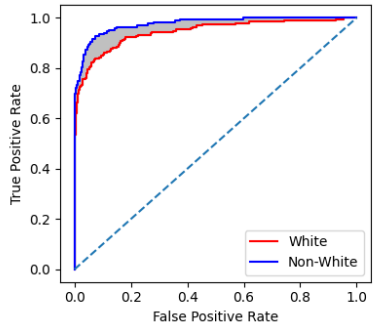


- Independent/Dependent Features Model - This model's ABROCA evaluation was similar to the Independent in that the Female/Male comparison was most fair (ABROCA 0.01), the Non-White/White comparison was higher (0.03) and the International/Non-International was only slightly higher. Like above, the model performed better for Female, Non-White, and International subgroups. However, the ABROCA scores in each case for this model are so low (<0.03) that the model could still be considered relatively fair. The differences could be driven by differences in balance across subgroups (same as above) or just random noise.

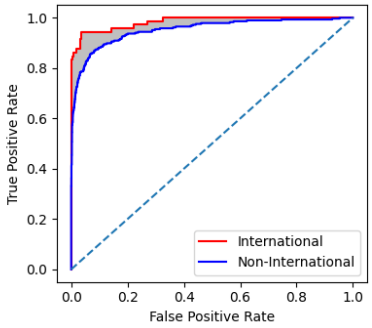
Gender Split
ABROCA: 0.00968



Race Split
ABROCA: 0.03194



Nationality Split
ABROCA: 0.03183

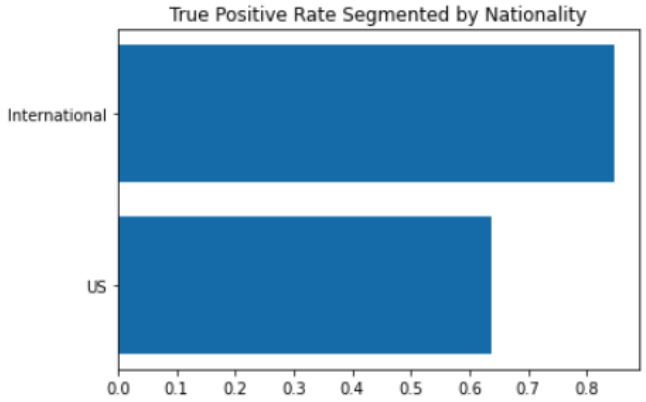


Other Model Bias Evaluation

To further investigate model fairness and bias, we examined demographic parity and equal opportunity for the dependent-independent model:

- **Demographic Parity**: The ratios of demographic segments selected by the model should be similar to the ratios of students actually at risk.
- **Equal Opportunity**: The true positive rate across demographic segments should be approximately equal.

What we found reinforced what we saw in the ABROCA plots, and the greatest disparity was that international students had a much higher true positive rate than American students (as seen below for the Independent/Dependent feature model).



Refer to Appendix F to see all demographic parity and equal opportunity plots.

Conclusions

Contributions of Our Work

Our work has made several contributions in the context of identifying academically at-risk undergraduate students:

Models and pipelines which could be used to identify students in need of support. Our final Independent Features Model (AUC 0.75) and our Independent/Dependent Features Model (AUC 0.96) both achieved respectable performance in predicting at-risk students, though the addition of Dependent features clearly improves predictive ability. The Independent Features Model allows making the predictions based upon static historical information and may enable setting up a support plan for specific students (or types of students) as soon as they are admitted. The Independent/Dependent Features Model allows identification and creation of a support plan while the student is "in-flight" at the university. In their current forms, with the DVC pipeline already established, these models also have the ability to (1) make at-risk predictions for new students using existing training data, (2) make at-risk predictions for new students based upon a different set of training data (such as future-year LARC data), or (3) to be the basis for further exploration based upon adjusting elements of our code and pipeline.

Identified, and ranked, features which could be viewed as early-warning indicators to be considered in policy development. When developing policy and programs regarding student support, it's important to identify key leverage points that may indicate signs of trouble and/or may be areas in which to invest time and resources. Our work has identified and ranked some key features which we believe contain important signals to which administrators should pay attention when they consider how to identify and support at-risk students. (Again, however, high importance does not definitively mean that the feature, individually, is a sign of risk. These features may be worth studying more). Important features include characteristics such as students from lower family income or educational levels, GPA means and/or variability, undeclared ("school career") status, and more.

Identification of less-obvious features which may merit further attention or exploration. While some obvious features (term GPA, family income, etc.) were identified as having high predictive value, some other less-obvious features were additionally identified as high importance. These include whether students come from specific countries (Korea, China, etc.) or are enrolled in specific schools (Music, Engineering, etc.). These student characteristics and associated outcomes may be worthy of further investigation into their relationship with academic success and, therefore, their individual predictive ability.

Avenues for Future Work / Limitations of Our Work

With a goal as broad as "predicting academically at-risk students", we were forced to make many decisions to make the data set and scope manageable. We built tools and drew conclusions within those confines, but along the way there were several possible improvements and unexplored avenues which may warrant further work.

Additional Feature Selection Approaches. We manually selected features based upon contextual knowledge, manually removed sparse features at a stated threshold, and performed automated feature pruning based upon permutation importance. We had additionally explored cosine similarity among features (Appendix B), though relied upon automated feature pruning to remove redundant features. Though we believe this to be a thorough approach, we're cognizant that (with more time) different decisions about feature inclusion may yield different results.

Deep dive into the term time series data. Each student has the same features collected for each term they attend. The term data for each student represents, in effect, a time series of each feature (some continuous, some categorical). While we leveraged the term data, and even looked at term-to-term GPA changes, there is an opportunity to build models that specifically take advantage of the time series for each student. In its simplest (though still complex) conceptual form, this might involve building sequence classification models for each feature and then using those outcomes as features in an at-risk student prediction model.

Further development of the term-by-term models. For efficiency reasons, we chose to further develop the single independent/dependent model (the one based upon all terms' worth of data) rather than further developing each of

the “first-two-terms” and “first-three-terms”, etc. models. While we believe that our chosen model was a proxy for the two dozen other models, it would be valuable to explore how each of those term models differ and whether/how the feature importance shifts as the term count increases.

Weighting of term data. Our evaluation approach of term data, in effect, equally weighted each term. However, one could make an argument that not all terms should be treated equally. A fall term will likely have a full load of concurrent classes. A summer term might have few classes with a more intense schedule, or might even have an internship/job in parallel. It’s unclear that they should be treated equally. Further, when training models based upon the term (a student’s first, second, etc.), one student’s third term may be summer while another student’s might be fall. It’s unclear that this is how terms should be aligned across students. While we believe that there is valuable information in the term data, it requires resolving some of these complicated issues.

Other predictive labels and definitions for an academically at-risk classification. While our use of “failing to graduate within 5 years” as the predictive label for identifying academically at-risk students is logical, other definitions may be worth further exploration. We briefly explored use of term GPAs below 2.0, though other GPA metrics, such as cumulative GPA or GPA in certain types of classes, may yield different results. Other definitions of at-risk might include low graduation class rank, “official” withdrawal from the university, or multiple consecutive class withdrawals (if all of these features become available). Further, rather than treating this as a classification problem, this could have been treated as a regression problem, such as predicting graduation GPA, graduation class rank, or time to graduate.

Further explore model fairness among subpopulations. Especially with globalization, academic environments and populations continue to diversify. We took advantage of the ABROCA approach to understand the fairness of our models among several subpopulations. However, there are many additional subpopulations, or other ways of segmenting the data, which may be valuable to explore. Further segmentation among students might include athletic participation and military status. Additionally, segmentation into multiple categories (e.g. football, track, tennis player, etc.) instead of binary participation segmentation (e.g. “athlete yes-vs-no”) may provide better insights. There are even further-reaching segmentations that may be valuable, such online versus residential classes and single students versus students with families, though LARC does not at present capture these characteristics.

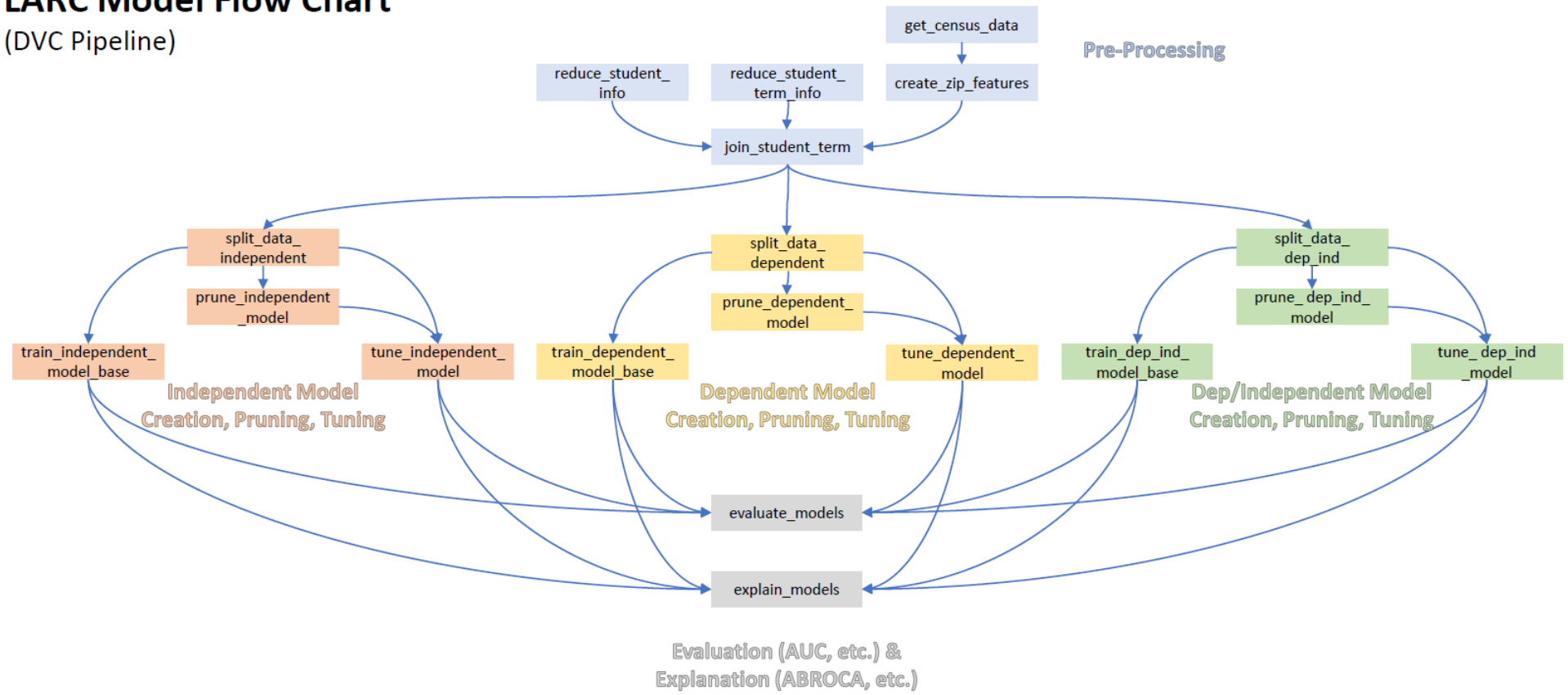
Applicability across institutions and time. Our current models were trained with Un of Michigan (Ann Arbor) data from a specific time period, so it may be valuable to evaluate how well these models perform when making predictions for students at other campuses or from different time periods. Additionally, because we built a pipeline, our models could be trained with data from other campuses or periods, if data is collected in the same manner.

Appendices

- Appendix A: Workflow / DVC Pipeline
- Appendix B: Pre-Processing Feature Evaluation
- Appendix C: AUC Performance Improvement Through Feature Pruning and Parameter Tuning
- Appendix D: SHAP Feature Importance Plots
- Appendix E: Confusion Matrices
- Appendix F: Additional Model Fairness Evaluation (Independent Dependent Model)

Appendix A

Workflow (DVC Pipeline)



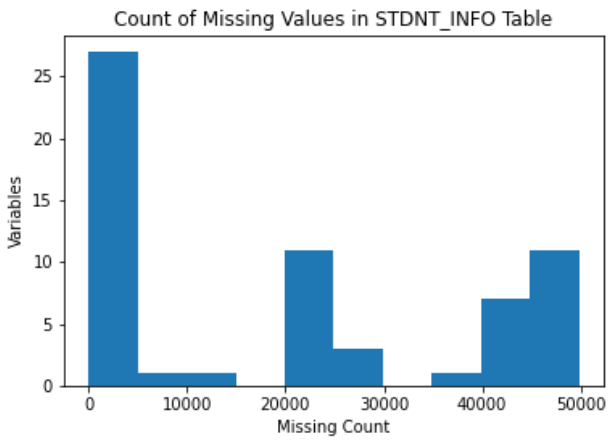
Stage	Description
get_census_data	Download school, economic, household, and internet availability data from census.gov API.
create_zip_features	Join and group census data based on zip codes.
reduce_student_info	Reduce data in student_info table based on years (2012 - 2015) and initially kept features.
reduce_student_term_info	Reduce data in student_term_info table based on years (2011 and beyond) and initially kept features.
join_student_term	Aggregate student_term_info by student. Join student_info, student_term_info, and census data.
split_data_independent	Perform imputation and one hot encoding. Split data into train, validation, and test sets
train_independent_model_base	Train baseline model for logistic regression, naive bayes, and random forest.
prune_independent_model	Perform feature tuning on baseline models.
tune_independent_model	Performs hyperparameter tuning on pruned models using grid search.
split_data_dependent	Perform imputation and one hot encoding. Split data into train, validation, and test sets
train_dependent_model_base	Train baseline model for logistic regression, naive bayes, and random forest.
prune_dependent_model	Perform feature tuning on baseline models.
tune_dependent_model	Performs hyperparameter tuning on pruned models using grid search.
split_data_dependent_independent	Perform imputation and one hot encoding. Split data into train, validation, and test sets
train_dependent_independent_model_base	Train baseline model for logistic regression, naive bayes, and random forest.
prune_dependent_independent_model	Perform feature tuning on baseline models.
tune_dependent_independent_model	Performs hyperparameter tuning on pruned models using grid search.
evaluate_models	Calculate roc auc, f1, recall, precision, and accuracy scores for all models. Plots roc curves and ABROCA for all models.
explain_models	Calculate permutation importance and generate partial dependence plots for all models.

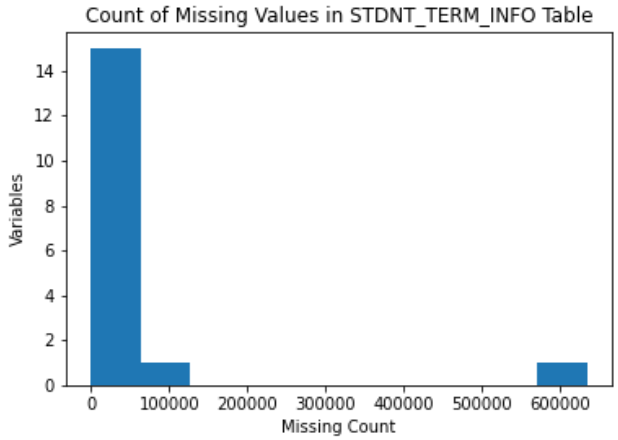
Appendix B

Pre-Processing Feature Evaluation

Sparse Variable Counts During Pre-Processing

Variables with a high number of missing values (STDNT_INFO: >10,000, STDNT_TERM_INFO: >100,000) were removed

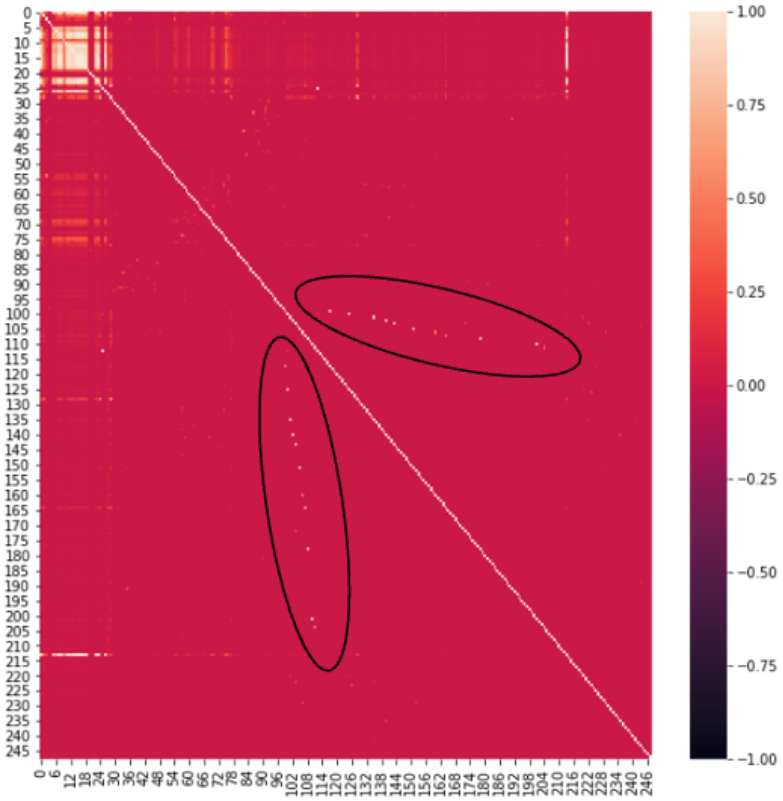




Feature Similarity Analysis During Pre-Processing

Feature similarity analysis (using cosine similarity) highlighted numerous highly similar features. The two larger categories of features that created high similarities were:

- Residency & Citizenship features (often the same country)
- ACT & other standardized test scores



Appendix C

AUC Performance Improvement Through Feature Pruning and Parameter Tuning

Validation Results

Model Type	Model	Baseline Performance	Post-Pruning Performance	Post-Tuning Performance
Independent	Naive Bayes	0.671661802	0.68101585	0.68101585
	Logistic Regression	0.736817689	0.750708861	0.748244751
	Random Forest	0.743028931	0.74732578	0.762109919
Dependent	Naive Bayes	0.865037206	0.896200029	0.896200029
	Logistic Regression	0.919162848	0.91960179	0.919341753
	Random Forest	0.9527929	0.957539094	0.95335354
Dependent-Independent	Naive Bayes	0.877510656	0.893648027	0.893648027
	Logistic Regression	0.926549976	0.930724088	0.92890279
	Random Forest	0.947754425	0.952826185	0.95659568

Test Results*

Model Type	Model	Baseline Performance	Post-Pruning Performance	Post-Tuning Performance
Independent	Naive Bayes	0.696753771	0.699534555	0.699534555
	Logistic Regression	0.736131009	0.732078175	0.73534067
	Random Forest	0.752573815	0.744472318	0.776184935
Dependent	Naive Bayes	0.881969946	0.906957903	0.906957903
	Logistic Regression	0.927558853	0.927484824	0.927592218
	Random Forest	0.95249572	0.952268941	0.955162854
Dependent-Independent	Naive Bayes	0.88633975	0.895869696	0.895869696
	Logistic Regression	0.931111208	0.93056381	0.931060118
	Random Forest	0.956831637	0.954713987	0.958285631

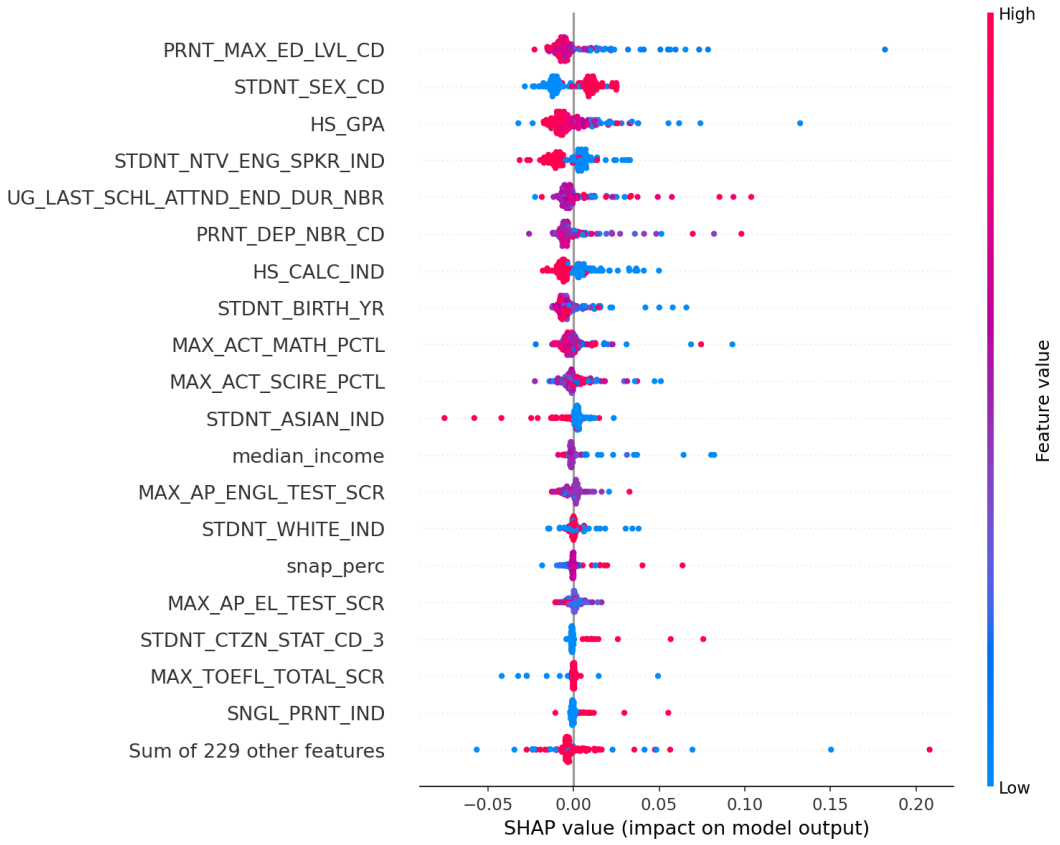
** Test results were not used to guide decisions during model iterations, and they were only generated at the completion of the project.*

Appendix D

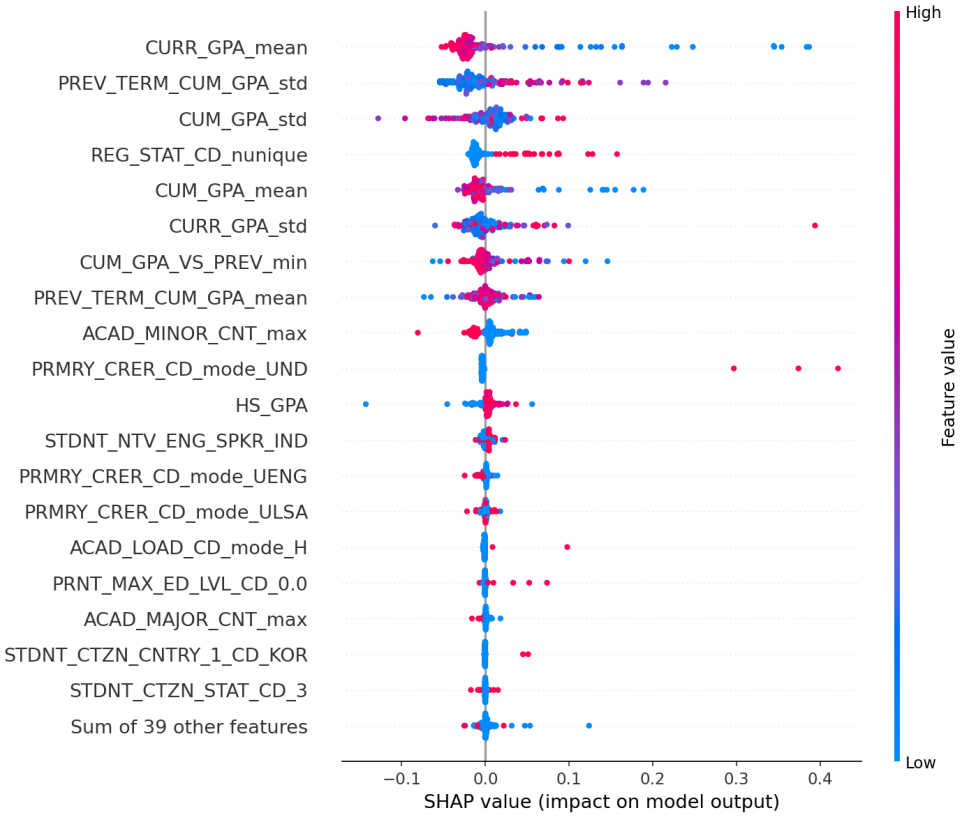
Shapley Feature Importance Plots

To further understand the feature importances, we used Shapley values with the Python SHAP library. Shapley values are the average expected marginal contributions of the features after all possible combinations of features have been considered. This allows us to get an idea of how exactly some features affect the model in addition to their overall importance. For instance, we can see that a high High School GPA is associated with a decreased chance of being an at-risk student.

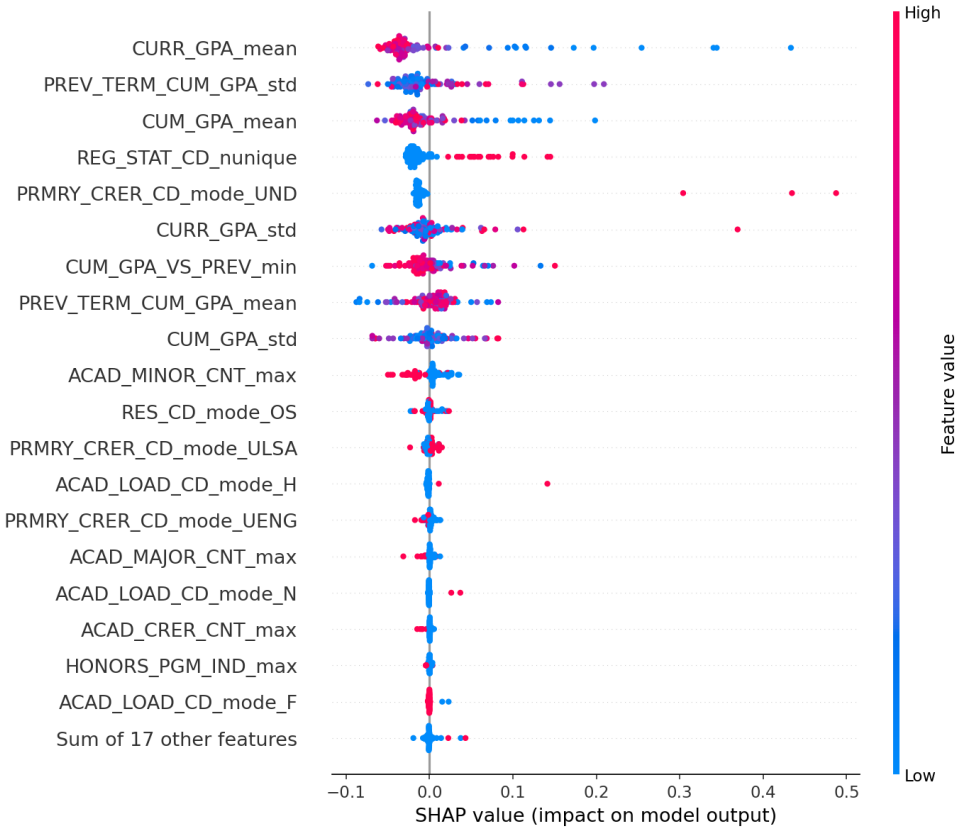
Independent Random Forest



Independent/Dependent Random Forest



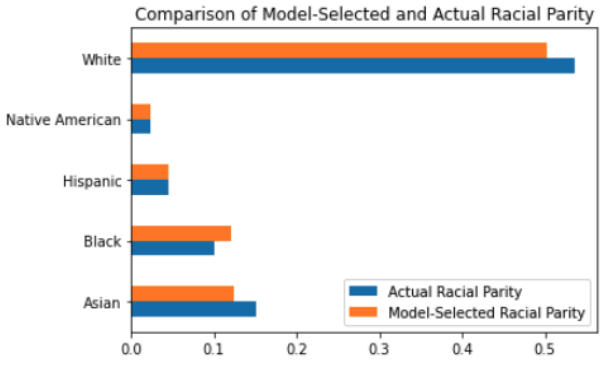
Dependent Model Random Forest

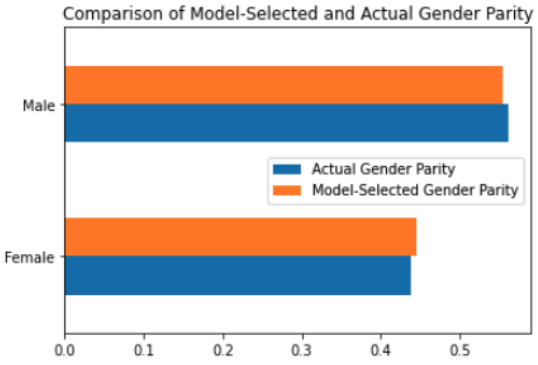


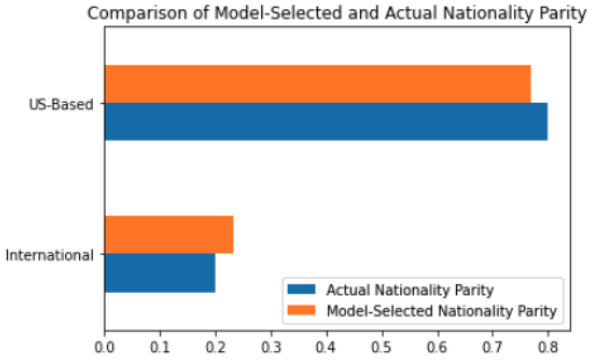
Appendix F

Additional Model Fairness Evaluation
(Independent Dependent Random Forest Classifier Model)

Demographic Parity







Equal Opportunity