SMU Data Science Review

Volume 1 | Number 2 Article 1

2018

# Machine Learning to Predict College Course Success

Anthony R.Y. Dalton
*Southern Methodist University*, adalton@mail.smu.edu

Justin Beer
*Southern Methodist University*, jbeer@mail.smu.edu

Sriharshasai Kommanapalli
*Southern Methodist University*, skommanapalli@smu.edu

James S. Lanich Ph.D.
*Educational Results Partnership*, jim@edresults.org

# Machine Learning to Predict College Course Success

Anthony Dalton[1], Justin Beer[1], Sriharshasai Kommanapalli[1],
James S. Lanich, Ph.D[2].

[1] Masters of Data Science Program at Southern Methodist University
6425 Boaz Lane
Dallas, TX 75205

[2] Educational Results Partnership
2300 N Street, Suite 3
Sacramento, CA 95816

**Abstract.** Herein we present an analysis of the predictive ability of machine learning on students' transfer-level course success in the California Community College system. The California Legislature passed assembly bill 705 in order to better place students in non-remedial coursework based on high school transcripts in order to increase college completion. Through utilizing machine learning methods on de-identified student high school transcript data, we create a predictive model for placing students in the correct initial college courses. The prediction is whether or not a student will be successful in transfer-level English and Mathematics courses. Industry knowledge is our guide to selecting variables to utilize in machine learning. The English dataset's best prediction model is the Random Forest with an accuracy of 65.7% along with better precision, recall, and support values than the other models. The Mathematics dataset's best prediction model is Logistic Regression at an accuracy of 67.1% and better precision, recall, and support values. The results indicate that students' subject cumulative grade point average is the best predictor for the English model and without subject cumulative grade point average is the best predictor for the Mathematics model. Overall, we built a predictive model that can be replicated and uniformly implemented which provides up to two times better accuracy as compared to current admission standards. Implementation of our model will benefit first time college students by accurate placement into courses they will be able to successfully complete.

## 1 Introduction

Equity in education is a topic that has long been researched, discussed, and funded. Tremendous efforts have been made to advance all students regardless of demographic characteristics. Research has shown that areas of handoff between education segments contain many forms of inequality in ethnicity, gender, and socioeconomic status. In particular, when first-time students seek admission into the California Community College system, and many other community college systems, they are given a placement test to assess whether they should enter into transfer-level courses or be placed in remedial courses. Unfortunately, there is no uniform model for assessing a

student's abilities as they enter college. Test selection and cut-scores are unique to each campus. Due to this, many college students suffer from being inaccurately placed in courses that do not count for credit or make the student feel like they are not college material. Remedial classes expend financial aid at the same rate as non-remedial coursework.

If an honors student applies to a community college and tests poorly, there is a chance that the student will have to take additional remedial courses before beginning their accredited college courses. Remedial courses are a sequence of up to eight preparatory courses to bring students up to "college-level" [1]. These courses do not count for credit and have the same fee associated as accredited coursework.

In the California Community College system, the largest education system in the nation, eight out of ten students are placed into remedial Math or English. African American and Hispanic students are placed at a rate of nine out of ten. Students who entered the California Community College System without starting in remedial coursework were 40% more likely to transfer to a four-year university [2]. The current placement test for most California Community Colleges is Accuplacer, from the College Board. In 2009, a research article published by College Board, shows that there is a very weak correlation between the Accuplacer test and success with a 'C' or better, with an average correlation of only 21% [3]. It is imperative that a model is created that will provide students with correct placement given their abilities.

In order to facilitate a better model for the California Community College system, the California legislature passed Assembly Bill 705 (Irwin) which was signed on October 13, 2017 by Governor Jerry Brown. The bill contains a section stating "this bill would require a community college district or college to maximize the probability that the student will enter and complete transfer-level coursework in English and Mathematics courses. In order to achieve this goal, one or more of the following: high school coursework, high school grades, and high school grade point average." [4]

This bill attempts to create a uniform standard by which the entire California Community College system can determine a student's likelihood to succeed in transfer-level coursework. The language in the bill states students should only be placed in remedial coursework if they are highly unlikely to succeed in college-level course work. This language is intentionally ambiguous to allow local control at the community colleges.

We will utilize de-identified student high school transcript data in order to create a model by which the California Community College system can accurately predict a student's success in transfer-level courses. An accurate model will limit the amount of money, financial aid, and time wasted by the student and college from incorrect placement. It is crucial for the analysis to be replicable, scalable, and reproducible to convince faculty and administration that the new method of placing students is a better alternative to current placement tests.

## 2 Related Work

The North Carolina Community Colleges have implemented a hierarchical structure of high school measures related to a 2.6 (C+) cumulative high school grade point average,

ACT, and/or SAT scores [4]. The RP Group and Education Results Partnership in California developed the first multiple measures model to predict students that were highly likely to succeed [5]. The models were built in RStudio with classification and regression trees.

The mission of the initiative was to utilize a disjunctive approach for students to take a standardized test as well as use predictive analytics for placement and for counselors and college administration to take the higher placement of both methods. The RP Group and Education Results Partnership approach was revolutionary in using predictive analytics for placement, but it isn't aggressive enough to satisfy AB 705. Due to the relative newness of attempting to predict college success and placement, there is a rarity of examples of implementation. The method that successfully accomplishes the goals of bill AB 705 will almost certainly set the precedent for college course placement methodology in the coming years.

## 3 Data Background and Information

The cohort that is included in this analysis is derived from the Cal-PASS Plus system of data. Cal-PASS Plus is an intersegmental data system that contains full transcript level data for all 114 California Community Colleges. Included are memorandums of understanding and data sharing with nearly 500 school districts that represent 80% of high school students in California. The dataset used in this analysis has all student identifiable fields removed and anonymized. From the student level records that exist in Cal-PASS Plus, records must meet four criteria. First, the Community College students must be enrolling in a credit section of English or Math for the first time, within 2012-2013, 2013-2014, 2014-2015, or 2015-2016. Second, students must have high school records that exist in Cal-PASS Plus. Third, students must have section-level records for 9$^{th}$, 10$^{th}$, and 11$^{th}$ grade. Fourth, students must not have any intersegmental key collisions.


**Fig. 1.** Illustration of 4 step criteria for finalized sample population.

# 4 Analysis

Analysis of data includes exploratory data analysis, finalizing input variables, creating the models, and analyzing the model metrics to find the best model. All code was peer reviewed as well as reviewed by James S. Lanich, Ph.D. for clarity, understanding, and evaluation. Dr. Lanich has over 20 years of experience in education serving as Assistant Superintendent of Curriculum and Instruction of Los Angeles Unified as well as serving on the National Assessment Governing Board. The variables that we are using to predict a college student's success are their high school grade point average, high school course grades, and high school coursework. While the bill provides that at least one of these classifiers needs to be used in the decision-making process, we will utilize as many as needed to create the most accurate model.

## 4.1 Exploratory Data Analysis

In order to maintain model validity, we ensured that our data meets the assumptions of the models we utilize. For the English dataset Fig. 2 demonstrates that the boxplots for the Overall GPA, Subject GPA, and Without Subject GPA are all left tailed. To confirm, Fig. 3 displays the distribution of the Overall GPA variable. Both figures display that the GPA data has a slight left tail.

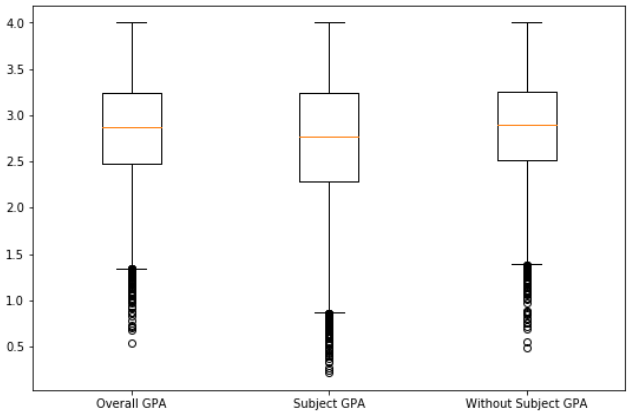


**Fig. 2.** Boxplots of English Variables.

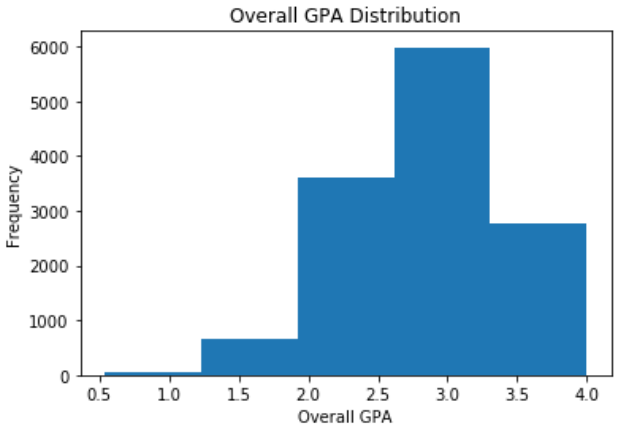


**Fig. 3.** Distribution of English Overall GPA.

Descriptive statistics were generated to check for any abnormalities. Fig. 4 displays that several math variables have a slight left tail. This is a common attribute between both the English and math variables. What the tail displays is that the mean grade points for these variables are all around 2.5 to 3.5. This will not affect the predictive ability to classify the data into a binary 1 or 0 category for the models utilized. The distribution is an interesting yet expected attribute which shows that high school teachers provide an average of a B and C for course grades. Fig. 5 confirms the tail for the Subject GPA variable.

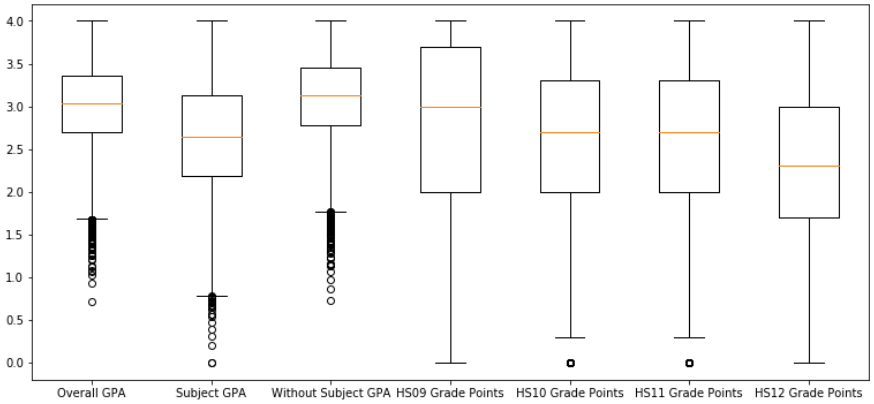


**Fig. 4.** Boxplots of Math Variable.

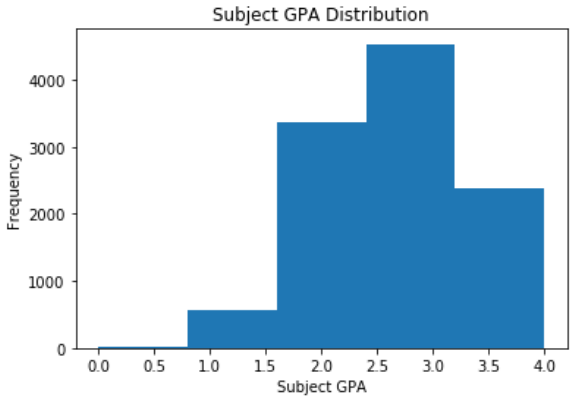


**Fig. 5.** Distribution of Math Subject GPA Variable.

### 4.2 Final English Variables

The response variable being classified and predicted is the "cc_00_course_success_id" which is either a 1 or 0. The value of 1 represents that the student was successful in their college course having passed the course with a C or better; a value of 0 represents students that were the logical opposite. Over 50 variables were considered as potential inputs to model selection. However, these were dropped one-by-one as we discovered they added no value to the models either domain wise or accuracy and precision wise. This was an iterative process to simplify the models as much as possible while retaining their accuracy and precision. Below are the variables that are utilized as input to the final models:

1. Overall Cumulative GPA: is as an average of all cumulative GPA's which a student has secured in all semesters and all the courses in an academic term.
2. Subject Cumulative GPA: is the grade point average of all grades a student has secured in English.
3. Without Subject Cumulative GPA: is the grade point average of all grades a student has secured in a semester or term outside of English.

### 4.3 Final Mathematics Variables

For the Mathematics dataset, the response variable is also "cc_00_course_success_id" and the model is attempting to predict its value of either a 1 or 0. Once more we utilized an iterative process to identify the most valuable indicators, having started with 180 in total, in order to obtain the simplest and most effective models. The variables that are utilized as input to the final models are:

1. Overall Cumulative GPA: is the average of all cumulative GPA's which a student has secured in all semesters and all the courses in an academic term.
2. Subject Cumulative GPA: is the grade point average of all grades a student has secured in the Math subject.
3. Without Subject Cumulative GPA: is the grade point average of all grades a student has secured in a semester or term without taking into account the Math subject.
4. hs_09_course_grade_points: Course grade points in 9th grade.
5. hs_10_course_grade_points: Course grade points in 10th grade.
6. hs_11_course_grade_points: Course grade points in 11th grade.
7. hs_12_course_grade_points: Course grade points in 12th grade.
8. hs_09_course_success_ind: Was the student successful in their 9th grade Math course? The value of 1 represents that this course was passed with a C or better and a value of 0 represents that the course was not passed.
9. hs_10_course_success_ind: Was the student successful in their 10th grade Math course? The value of 1 represents that this course was passed with a C or better and a value of 0 represents that the course was not passed.
10. hs_11_course_success_ind: Was the student successful in their 11th grade Math course? The value of 1 represents that this course was passed with a C or better and a value of 0 represents that the course was not passed.
11. hs_12_course_success_ind: Was the student successful in their 11th grade Math course? The value of 1 represents that this course was passed with a C or better and a value of 0 represents that the course was not passed.

### 4.4 Methodology for Predictive Models

Both supervised and unsupervised learning techniques are employed to produce the most accurate model for prediction. The algorithms used are logistic regression, classification and decision trees, adaboost with decision trees, support vector machines, artificial neural networking, naïve Bayes, and random forest. In order to train and test the models we used the ShuffleSplit from the Python package Sklearn. The data was iterated over 10 times and 80% of the data was used for training and 20% of the data for testing. After building and testing the models we obtained the results as defined by accuracy, precision, recall, and f-score. We chose the best model based upon these metrics. Once the best model is identified we further validate it against previous years of data to ensure accuracy of model fit and protection against over-fitting.

### 4.3 Summary of Findings for English Dataset

For the English dataset, 32,092 records were used in the training and 9,122 records were withheld for further testing. The following are the models we evaluated with their explanations, followed by a table of results, and noteworthy figures.

**Logistic regression**. A logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of the presence of our dependent variable: cc_00_course_success_ind.

$$\text{Logit}(p) = a + b*\text{coursework} + c*\text{grades} + d*\text{GPA}\ldots\ldots$$

Where b, c, and d are the weights of the logistic regression associated with the independent variables, the higher the weight, the higher the influence of that variable. We found that the most influential predictor for the logistic regression model is the Without Subject Cumulative GPA with a weighting of 0.83, then the Subject Cumulative GPA with a weighting of 0.82, and finally the Overall Cumulative GPA with a weighting of -0.41.

**Naive Bayes.** The Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature, the features are considered to contribute independently to the probability of cc_00_course_success_ind to be classified as one of its outcomes.

**Decision Tree.** We evaluated Classification and Regression Trees (CART) with rules for splitting data at each node based on the values of the independent variables, stopping rules for when a branch is terminal, and a prediction for "cc_00_course_success_ind" at each terminal node. The best CART model had the criterion=entropy, max depth=20, minimum leaf samples=5, minimum samples split=5, and maximum leaf nodes=62.

**Adaboost with Decision Tree.** Including the Adaboost function with the use of CART can improve the overall model performance due to the meta-weighting provided by Adaboost. The best model had the algorithm = SAMME.

**Random Forest.** We used the random forest classifier which uses a multitude of decision trees. The best results were obtained using criterion = Gini, minimum samples per leaf = 0.1, and minimum samples per split = 0.2. Fig. 6 shows the AUC of Precision Recall curve having a value of 79.6%. The weights for the Random Forest are Subject Cumulative GPA = 0.56, Overall Cumulative GPA = 0.15, and Without Subject Cumulative GPA = 0.13.

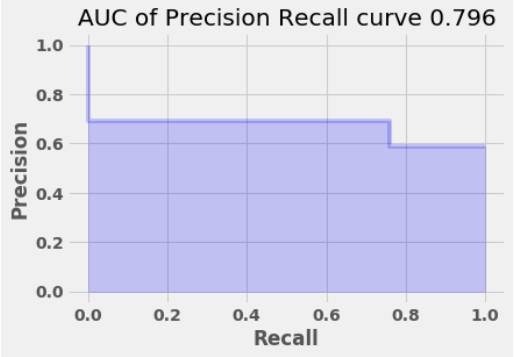


**Fig. 6.** AUC Precision-Recall Curve.

**Support Vector machine**s. SVM considers each record of dependent and independent variables as (x, y). Where x is a p-dimensional linear vector, p is the number of independent variables, and y is the value of the dependent variable "cc_00_course_success_ind". We will utilize linear SVM on these data to find a hyperplane that can linearly separate the data for the best classification.

**Artificial Neural Network (ANN)**. The neural network contains three layers. They are an initial layer of "tanh" activation, a secondary hidden layer with "relu" activation, and a third hidden layer with "sigmoid" activation. We used the loss = binary_crossentropy to minimize miss-classifications and the optimizer = "adam". Fig. 7 shows an average precision-recall score of 76%.

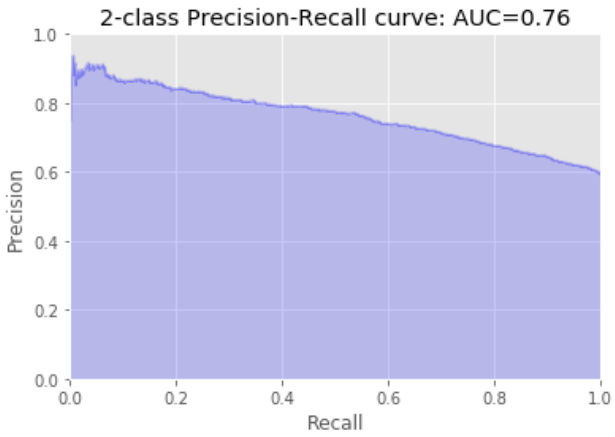


**Fig. 7.** ANN Precision-Recall Curve.

The following table is a summary of the model metrics obtained for the English dataset.

**Table 1.** English dataset models and corresponding metrics. The Support is defined as a count of the 0-Fail, 1-Pass. The Precision, Recall, and F-Score are in the percent format 0-Fail, 1-Pass.

| Model | Avg. Accuracy | Precision | Recall | F-Score | Support |
|---|---|---|---|---|---|
| Logistic Regression | 66.3 | 62.5, 66.1 | 41.8, 81.9 | 50.0, 73.2 | 1450, 2150 |
| Naïve Bayes | 65.1 | 59.2, 68.9 | 54.3, 72.9 | 56.6, 70.8 | 1509, 2091 |
| Decision Tree | 64.6 | 56.1, 70.1 | 55.2-72.0 | 55.7, 70.5 | 1450, 2150 |
| Adaboost w/ Decision Tree | 65.4 | 57.9, 69.6 | 51.7, 74.7 | 54.6, 72.0 | 1450, 2150 |
| Random Forest | 65.7 | 60.0, 69.0 | 51.4, 76.0 | 55.4, 72.4 | 1485, 2115 |
| Support Vector Machines | 65.8 | | | | |
| Artificial Neural Network | 66.0 | 59.3, 68.8 | 48.8, 77.1 | 53.5, 72.0 | 1458, 2142 |

Table 1 demonstrates that the best performing model is the Random Forest Model as it has the best mix of accuracy, precision, and recall. By testing this Random Forest model on our hold out set of 9,122 records, we achieved the below results. These validate the model parameters generated for the Random Forest model in Table 1.

**Table 2.** Results of using our created Random Forest model on the 9,122 record holdout set.

| Model | Avg. Accuracy | Precision | Recall | F-Score | Support |
|---|---|---|---|---|---|
| Random Forest | 67.9 | 46.4, 79.4 | 54.5, 73.5 | 50.1, 76.3 | 2699, 6422 |

**Table 3.** Confusion Matrix for Random Forest model used on the 9,122 record holdout set.

| | Actual Fail | Actual Pass |
|---|---|---|
| Predicted Fail | 1471 | 1228 |
| Predicted Pass | 1699 | 4723 |

### 4.4 Summary of Findings for Mathematics Dataset

For the Mathematics dataset, we used 8,690 records to train the models and held aside 2,172 records for testing. Notable findings are that for Logistic Regression the most influential predictor is the Without Subject GPA. For the Decision Tree, the best model had the criterion = entropy, max depth = 20, minimum leaf samples = 5, minimum samples split = 5, and maximum leaf nodes = 62. For the Random Forest model, the best results were obtained using criterion = entropy, maximum depth = 10, minimum samples per leaf = 5, minimum samples per split = 5, and maximum leaf nodes = 62. Random Forest also had an AUC of Precision Recall curve with a value of 82%. The Artificial Neural Network had an average precision-recall score of 79%.

**Table 4.** Mathematics dataset models and corresponding metrics. The Support is defined as a count of the 0-Fail, 1-Pass. The Precision, Recall, and F-Score are in the percent format 0-Fail, 1-Pass.

| Model | Avg. Accuracy | Precision | Recall | F-Score | Support |
|---|---|---|---|---|---|
| Logistic Regression | 67.1 | 60.9, 68.2 | 37.1, 85.0 | 46.1, 75.6 | 841, 1332 |
| Naïve Bayes | 64.8 | 56.6, 70.7 | 54.4, 72.5 | 55.5, 71.6 | 864, 1309 |
| Decision Tree | 66.5 | 57.9, 69.2 | 37.6, 83.7 | 45.6, 75.8 | 812, 1361 |
| Adaboost w/ Decision Tree | 65.7 | 55.9, 69.2 | 39.2, 81.6 | 45.8, 74.8 | 812, 1361 |
| Random Forest | 66.2 | 59.8, 70.5 | 42.4, 82.8 | 49.6, 76.2 | 816, 1357 |
| Support Vector Machines | 67.4 | | | | |
| Artificial Neural Network | 63.5 | 57.1, 70.1 | 47.7, 77.4 | 52.0, 73.6 | 841, 1332 |

### 4.5 Experiment on English non-transfer student records

After we trained and tested our models, we applied the Random Forest model to a set of English non-transfer students. We utilized a non-transfer student set consisting of 102,372 records, which had no role in creating the transfer models in section 4.3. The goal of this process was to assess our models mimicking an example of real-world performance. The results indicate which students would pass transfer level courses and should be allowed to enroll in a transfer-level college English course. Compared to the current testing standard, the results show that there is a 200% increase in students that can be enrolled in and pass a transfer-level college English course. African American and Hispanic students benefit the most, with an increase of 1,965 and 24,962 students respectively, and are projected to pass non-remedial coursework.

Previous research shows the most powerful and highest correlated transcript data to college course success is GPA [7]. Our research is consistent as we find that the Subject Cumulative GPA is the most influential predictor. While the usage of demographic characteristics helps increase accuracy, students should not be judged by demographics and were therefore not included in the final model. Fig. 8 and Fig. 9 show the benefits of using our machine learning models.

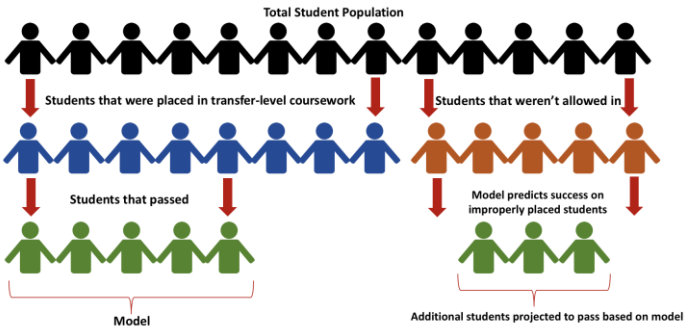


**Fig. 8.** New models provide greater opportunities for students.

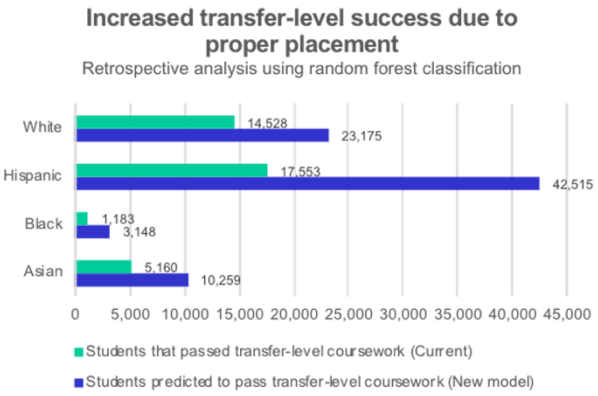


**Fig. 9.** Increased transfer-level success.

# 5 Ethics

## 5.1 Student Level Data

With student level records, FERPA compliance is critical. No identifiable information is provided nor included in this analysis. Directory information to be avoided in order for compliance are as follows:

1. Name, address, telephone number, date of birth, and grade level.
2. Participation in sports or other recognized activities.
3. Height and weight of athletes.
4. Degrees, honors, and awards received.
5. Most recent school attended.

Since High School GPA is the primary predictor for our models, there is a concern that grades may not be equated across districts or across years as teachers can grant grades that they see fit. However, there have been numerous studies that prove that grade inflation is largely a myth. Our model has been developed using all available relevant data. If non-normalized grades exist, they are also included in the data and tracked into college outcomes. “Good” grades from bad districts are often declared undeserved due to demographic make-up without any evidence to prove it [6]. Our goal, is to use the largest universe of data to provide a statewide model that can be used at scale for all students in the California Community College system. If models were developed for specific districts bias would be introduced that would be as harmful as the current standardized testing.

### 5.1 Confounding Factors

While assembly bill 705 attempts to base college admissions and placement into remedial courses on the previous performance of the student applying, there are manifold confounding factors that will affect the students' course success after admission. These confounding factors may be circumstantial for the individual admitted. The student may experience hardship in their lives that causes them to lose focus on their coursework yielding negative performance in the remedial courses. Also, an individual who is highly likely to succeed may end up failing due to their inability to cope with the stress associated with transfer-level courses. The confounding factors center on the individual, their circumstances, and their response to events. Unfortunately, there is not a way to predict whether an individual will experience a circumstantially difficult time during their attendance. These are the situational unknowns that colleges will have to find a separate way to deal with and support while the person is in attendance.

## 6 Conclusions

We find that the best model for predicting the success, whether a student will pass their college course with a C or better, of a student differs between our English and Math datasets. The English datasets best prediction model is the Random Forest with an accuracy of 65.7% and better precision, recall, and support values than the other models. The Mathematics datasets best prediction model is Logistic Regression at an accuracy of 67.1% and likewise better precision, recall, and support values than the other models.

Interestingly, all of the models we utilized outperformed the current standards by approximately 200%. We have proven that machine learning models are more capable and accurate than current placement standards. Using these models will increase the accuracy and precision of student placement in either remedial or transfer-level coursework. Employing the models will decrease time and funding wasted by both the student and college. Finally, these models provide a chance for more students to be able to stand on their own abilities throughout the admission process.

The benefit is that an equitable throughput increase can happen without needing to drastically overhaul practices between high school and college. To utilize a predictive model requires trust between institutions. A college has to assume that students are coming in prepared if their body of work over four years demonstrates an average GPA. With this trust, positive conversations can occur related to co-designing curriculum as opposed to blame and finger pointing.

By 2040, the workforce will be majority-minority. Without more equitable placement practices, we will not have enough middle-skilled or high-skilled employee to fill the workforce positions needed.

There is still much work to be accomplished in predicting a student's success in college. Our research is general and a broad look into the California Community College system as a whole and does not account for individual colleges or even the high schools previously attended by the students. However, the new models satisfy the

requirements of assembly bill 705 by utilizing high school grades and high school GPA. All students greatly benefit but are still in the category of highly likely to succeed. To start the piloting process of the new methodology, we will perform a randomized control trial to validate our findings with a willing college and feeder high school district.

## References


1. Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. Int. J. Digit. Libr. 1 (1997) 108–121
2. O. Rodriguez, "Remedial Education in California's Colleges and Universities," Public Policy Institute of California, October 2017. [Online]. Available: http://www.ppic.org/publication/remedial-education-in-californias-colleges-and-universities/. [Accessed 16 Februar 2018].
3. K. Mattern and S. Packman, "Predictive Validity of ACCUPLACER® Scores for Course Placement: A Meta-Analysis," College Board, 2009.
4. van Leeuwen, J. (ed.): Computer Science Today. Recent Trends and Developments. Lecture Notes in Computer Science, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York (1995)
5. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
6. A. Kohn, "The Dangerous Myth of Grade Inflation," 8 November 2002. [Online]. Available: https://www.alfiekohn.org/article/dangerous-myth-grade-inflation/. [Accessed 31 March 2018].
7. C. Belfield and P. Crosta, "Predicting Success in College: The Importance of Placement Tests and High School Transcripts," February 2012. [Online]. Available: https://ccrc.tc.columbia.edu/media/k2/attachments/predicting-success-placement-tests-transcripts.pdf. [Accessed 2 April 2018].