Predicting Student Success to Ensure Equity for Students

Jasmin Cornejo

B.S. Candidate, Department of Computer Science,, California State University Stanislaus, 1 University Circle, Turlock, CA 95382

Received 9 May, 2020; accepted 15 May, 2020

Abstract

In this paper we use machine learning algorithms to create predictive models of student success on the Writing Proficiency Screening Test (WPST) at California State University, Stanislaus. The data used included academic history, SES (socioeconomic standing), and demographic for the 3357 students who were tested in 2009-2011 at Stan State. The goal of this research project was to better identify "at risk" scholars before taking the WPST. The WPST is a graduation requirement for all students and to fail this exam may mean another semester in school. Identifying at risk scholars with predictive models may be the key in reducing cost and length of time in school.

Introduction

In 2019 college students across California will have to undergo some form of an English proficiency exam before the completion of their undergraduate degree. At California State University (CSU) Stanislaus, this exam is called the Writing Proficiency Screening Test (WPST). According to the CSU Stanislaus WPST website, "The Writing Proficiency Screening Test (WPST) assesses your readiness for Writing Proficiency (WP) courses, which are writing intensive courses designed to teach discipline-specific writing conventions." ("Writing Proficiency", n.d.) It is important to note that WP courses are necessary for graduating in a timely manner and failing the WPST can mean waiting another semester before taking these vital classes. Moreover, it is common knowledge that student debt is on the rise and the increased cost of attending school for an extended amount of time is money that could be used for more useful purposes. Hence, it is important for schools to identify "at risk" scholars who could benefit from extra help.

At CSU Stanislaus, 67% of students are first generation students, which means that neither of the students' parents attended college. ("Stan State Enrollment", 2018) It is well documented that first-generation students are at higher risk of prolonging school and dropping out. According to National Center for Education Statistics (NCES), it has been found that first-generation students drop out of college at a proportionally higher rate than students whose parents graduated from college. (Bennett, Chen, & Cataldi, 2018) Furthermore, it has been shown that first-generation students generally stay up to 6 years in college, with the NCES reporting that 60% of students that were first-time, full-time undergraduates in fall 2010 graduated 6 years later, or 150% of the standard timeframe. ("The NCES Fast Facts", 2018)

It is vital for universities to do everything they can for students to stay on the path to success and graduate in a timely manner. With that in mind, many people are looking to technology for solutions. All colleges and universities in the United States are required to collect data on their students, which means there is an abundance of data waiting to be explored. Computers make it faster and more cost-effective than ever to analyze large sets of data. This area of Computer Science has become known as Data Mining.

Data mining uses machine learning algorithms to create predictive models that may explain relationships between demographic and income, or any other attributes associated with a big set of data. Its predictive power could be used to identify "at risk" students, which universities could then extend assistance to.

Background

Previous studies have already shown that it is possible to create an accurate model to predict student success. One study completed at Iowa State University found a model that could perfectly predict passing students from a training data set (0% error) and successfully forecast a test set (8% error). (Vu, 2016) The study noted that there was not a model that could successfully predict student scores on their English proficiency test. However, there were factors that could accurately predict highest and lowest essay scores. (Vu, 2016) Another study done at University of New Mexico found that they could correctly predict the future progress of a student using first semester grades with a margin of error of 0.16. (Slim, Heileman, Kozlick, & Abdallah, 2014) These studies demonstrate that we can reliably use students' past performance statistics as a measure for future progress.

Other studies have shown that data mining could be used to target "at-risk" students more precisely. For example, a study conducted in Singapore concluded that data mining increased the efficiency of selecting potentially weak students for remedial classes, which they claimed reduced the burden on both students and teachers. (Ma, Liu, Wong, Yu, P., & Lee, 2000) Another study found that with machine learning, the problem of student retention could be identified as early as the third week of a semester with 97% accuracy, as well as identifying students less likely to achieve a passing grade as early as the fourth week with an accuracy of 97.2%. (Gray, C. C., & Perkins, D, 2019) This shows that not only is it possible to accurately predict the success of students, but that with machine learning it can be done quickly and efficiently. This could allow for intervention programs to be put into place to help struggling students before it is too late to change their course in a semester.

Lastly, it has been shown that it is possible to identify factors that help predict student success. For instance, a study in Brazil found that grades and absences were the most relevant factors for predicting the end of the year academic outcomes of their students. In addition, demographic factors revealed that neighborhood, school, and age were also potential indicators of either student success or failure.(Fernandes, Holanda, Victorino, Borges, Carvalho, & Erven, 2019) Moreover, another study done with data from approximately 200,000 high school students, located in two separate school districts, found that 8th grade GPA was a highly predictive attribute of on-time high school graduation. (Lakkaraju, et al.,2015) This further demonstrates that machine learning algorithms can identify major factors in predicting student outcomes.

Thesis and Rationale

By using modern machine learning and data mining techniques, institutions of higher education can greatly increase retention rates, especially among first-generation scholars. There are some questions that need to be addressed to more efficiently meet this goal, which include: With academic, financial, and demographic data of students, can we make a model that can accurately predict student outcomes on the WPST? Which attributes hold the most weight in accurately predicting student outcomes on the WPST?

With something as significant as education, it cannot be understated how important it is to find and rectify negative patterns quickly so that a student has the tools they need to succeed. This will help ensure equity for all students and hopefully encourage others to seek higher education that may have previously seen such a pursuit as out of their reach. I believe that machine learning is the best and most efficient way to accomplish this task. The advent of machine learning techniques allows us to identify patterns, as well as possible outcomes, far more quickly and with greater accuracy than ever before.

Method

I received data from the Computer Science department of California State University, Stanislaus that was already anonymized. It contained information on 3,357 students. The data contained information on students' academic history, SES (socioeconomic standing), demographic and WPST score. All data received was from the years 2009-2011. Upon receiving and before analyzing the data, I preprocessed the data, to address issues such as missing values, outliers, and differences in data granularity. Since, we were trying to model the first-time test taking 579 instances were removed where the instances were the second or more time attempting the WPST.

During the preprocessing of the data it was also found that transfer students were missing data associated with the ACT (American College Test) and SAT(Scholastic Aptitude Test). So, to address this issue the data was split into two groups. The Community College transfer group (CC group) and the students who entered Stanislaus State straight from high school (HS group). The CC group contained a total of 1733 instances. While the HS group contained a total of 1044 instances. Of the 1733 CC group instances 314 were failed and 1419 were passed (CC group, 18% fail,82% pass). Among the HS group of 1044 instances 162 were failed and 992 were passed (HS group, 16% fails, 84% pass).

The data sets were unbalanced, meaning that there were more instances of pass than fail. Imbalanced data sets tend to impact the effectiveness of the algorithms to predict student outcomes. Moreover, the algorithm might be overwhelmed by the majority class and not be able to discern what makes a student fail.(Xu-Ying ,Jianxin Wu & Zhi-Hua Zhou, 2009) To combat this we implemented the tactic of under sampling the majority class, which means that we randomly selected pass instances and randomly left some out so that the data set would be even. To illustrate the HS group had 162 fail and 992 pass instances. Therefore, under sampling the majority class would mean that we randomly select 162 instances out of the 992 pass instances. Although we under sampled the majority class for training purposes, we tested the models on the real-world ratio by adding pass instances to the test sets. By randomly adding pass instances, we were able to obtain the CC group ratio of 18% fail/82% pass and the HS group ratio 16% fails/84% pass for the testing sets.

In addition, I explored methods to filter out redundant attributes, such as the Weka built in wrapper function and looking at the correlation of attributes. Redundant attributes were attributes that told the same information. For example, the English as a Second Language (ESL) and language spoken. They basically say the same information and thus don't add information for modeling purposes. I then used Weka to run machine learning algorithms on the data with the selected attributes.

Weka is an open-source program that contains a collection of machine learning algorithms used for data mining tasks. It also contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.(Eibe, Mark, & Ian, 2016) After preprocessing the data was run on the Weka algorithms J48, Random Forest, Naive Bayes, and SMO.

Using the 10-fold cross-evaluations a standard machine learning technique to avoid overfitting the data. Overfitting produces a model that corresponds too closely or precisely with a particular data set. Models that are overfitted tend to fail at making predictions about future data.

The general procedure for cross 10 evaluation is as follows:

1) Shuffle the dataset randomly.
2) Split the dataset into k (in our case k = 10) groups.
3) For each unique group:
    a) Take the group as a hold out or test data set.
    b) Take the remaining groups as a training data set.
    c) Fit a model on the training set and evaluate it on the test set.
    d) Retain the evaluation score and discard the model.
4) Summarize the skill of the model using the sample of model evaluation scores (Brownlee, 2018)

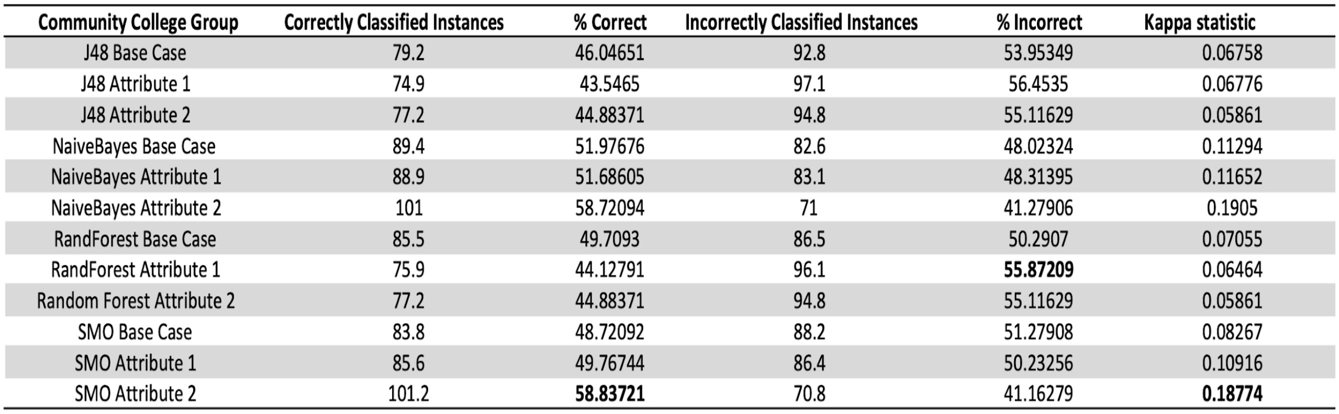
Lastly, I compared the different models by looking at the averages of the 10 models that were produced by cross 10-fold evaluation. I also considered their kappa statistics which is a statistic that looks to see if the model correctly classifies both cases of pass or fail.

Results

The results for the models for Highschool and CC group are summarized below in the tables.

High School Group	Correctly Classified Instances	% Correct	Incorrectly Classified Instances	% Incorrectly	Kappa statistic
J48 Base Case	74.1	**74.1**	25.9	25.9	0.19175
J48 Attribute 1	59	59	41	41	0.10942
J48 Attribute 2	61.1	61.1	38.9	38.9	0.19203
NaiveBayes Base Case	67.7	67.7	32.3	32.3	**0.22189**
NaiveBayes Attribute 1	63.9	63.9	36.1	36.1	0.2081
NaiveBayes Attribute 2	61.9	61.9	38.1	38.1	0.21868
RandForest Base Case	64.1	64.1	35.9	35.9	0.19475
RandForest Attribute 1	63.8	63.8	36.2	36.2	0.15797
Random Forest Attribute 2	60	60	40	40	0.14271
SMO Base Case	57	57	43	**43**	0.09668
SMO Attribute 1	60	60	40	40	0.17211
SMO Attribute 2	57	57	43	**43**	0.16826

(High School Group Table:
Base Case Attributes: 51, SexCode, RaceEthnicity, EthnicCode, CitizenCode, CountryCCode, HSGradYear, InstOriginCode, InstName, EnrollStatusCode, StudentLevelCode, College, Department, Major, DegObjCode, EmphasisCode, TransferUnitsEarned, TransferGpa, CampusGpa, TotalUnitsEarned, TotalGPa, EptCode, ElmCode, EOPCode, CriticalThinkingCourse, EnglishCompositionCourse, MathQuantReasoningCourse, OralCommunicationCourse, SAT1WritingScore, ACTWritingScore, TotalUnitsAttempted, HSGPA, ACTEnglishScore, ACTMathScore, ACTReadingScore, ACTScienceReasoning, ACTCompositeScore, ELMTotalScore, EPTEssayScore, EPTReadingScore, EPTCompositionScore, EPTTotalScore, SAT1VerbalScore, SAT1MathScore, SAT1CompositeScore, ESL, DB, LEVEL, LANG, SPEC, EFC, LOW_INCOME
Attributes 1 HS Group: 21, RaceEthnicity, EthnicCode, CitizenCode, CountryCCode, InstOriginCode, EnrollStatusCode, CampusGpa, TotalGPa, EptCode, EOPCode, HSGPA, ACTEnglishScore, ACTReadingScore, EPTTotalScore, SAT1VerbalScore, SAT1MathScore, SAT1CompositeScore, ESL, LANG, EFC, LOW_INCOME
Attribute 2 HS Group J48: 6, SexCode, EmphasisCode, EptCode, ElmCode, ACTReadingScore, Date of Birth
Attribute 2 HS Group Naive Bayes:4, RaceEthnicity, Major, EptCode, SAT1WritingScore
Attribute 2 HS Group RandomForest: 3, RaceEthnicity, EthnicCode, SPEC
Attribute 2 HS Group SMO: 7, RaceEthnicity, CitizenCode, EnrollStatusCode, Major, EptCode, EOPCode, SAT1MathScore)
*Bolded numbers are to show the highest number in that category

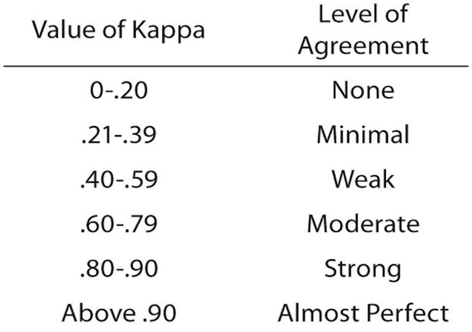


Community College Group	Correctly Classified Instances	% Correct	Incorrectly Classified Instances	% Incorrect	Kappa statistic
J48 Base Case	79.2	46.04651	92.8	53.95349	0.06758
J48 Attribute 1	74.9	43.5465	97.1	56.4535	0.06776
J48 Attribute 2	77.2	44.88371	94.8	55.11629	0.05861
NaiveBayes Base Case	89.4	51.97676	82.6	48.02324	0.11294
NaiveBayes Attribute 1	88.9	51.68605	83.1	48.31395	0.11652
NaiveBayes Attribute 2	101	58.72094	71	41.27906	0.1905
RandForest Base Case	85.5	49.7093	86.5	50.2907	0.07055
RandForest Attribute 1	75.9	44.12791	96.1	**55.87209**	0.06464
Random Forest Attribute 2	77.2	44.88371	94.8	55.11629	0.05861
SMO Base Case	83.8	48.72092	88.2	51.27908	0.08267
SMO Attribute 1	85.6	49.76744	86.4	50.23256	0.10916
SMO Attribute 2	101.2	**58.83721**	70.8	41.16279	**0.18774**

(Community College Group Table:
Base Case Attributes:36, SexCode, RaceEthnicity, EthnicCode, CitizenCode, CountryCCode, HSGradYear, InstOriginCode, InstName, EnrollStatusCode, StudentLevelCode, College, Department, Major, DegObjCode, EmphasisCode, TransferUnitsEarned, TransferGpa, CampusGpa, TotalUnitsEarned, TotalGPa, EptCode, ElmCode, EOPCode, CriticalThinkingCourse, EnglishCompositionCourse, MathQuantReasoningCourse, OralCommunicationCourse, TotalUnitsAttempted, ESL, DB, LEVEL, LANG, FRESH, SPEC, EFC, LOW_INCOME
Attribute 1: 17, RaceEthnicity, EthnicCode, CitizenCode, CountryCCode, EnrollStatusCode, StudentLevelCode, CampusGpa, TotalUnitsEarned, TotalGPa, ElmCode, MathQuantReasoningCourse, TotalUnitsAttempted, ESL, LEVEL, LANG, EFC, LOW_INCOME
Attribute 2 J48: 3, RaceEthnicity, EthnicCode, StudentLevelCode
Attribute 2 Naive Bayes:6, SexCode, EthnicCode, OralCommunicationCourse, ESL, LEVEL, LOW_INCOME
Attribute 2 Random Forest: 7, SexCode, RaceEthnicity, EthnicCode, StudentLevelCode, ElmCode, ESL, LEVEL
Attribute 2 SMO: 5, Department, DegObjCode, ElmCode, ESL, Pass or Fail, LEVEL)
*Bolded numbers are to show the highest number in that category

Discussion

To discuss the results of the models we must first understand Cohen's kappa Statistic. Cohen's kappa statistic measures how the test set of data was classified by the model. It considers that the amount of correctly classified instances can be high, but that the model itself may be weak. When looking at the data the following table shows the breakdown of the kappa statistic.



Value of Kappa	Level of Agreement
0-.20	None
.21-.39	Minimal
.40-.59	Weak
.60-.79	Moderate
.80-.90	Strong
Above .90	Almost Perfect

(Kappa Statistic table: McHugh, M. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica,22*(3), 276-282.)

As we can see from the above table the level of agreement with the original test set may be high or low.

If we look at the HS Group Table, we can see that the kappa statistic for the models across the board never goes over 0.25. Although, on average the models classified about 50% of the training set correctly. For the algorithm J48 the highest correctly classified was 74.1%. which was the base case. Some might think that this means the model was doing well. However, the kappa statistic for the J48 base case was 0.19175 which means that the model was not great at classifying fail. It can be explained like this. If our current test set is 80% pass and 20% fail, then even if it classifies every instance as pass the model will still be correctly classifying 80% of the data set. However, this is not a great model and will have a very low kappa statistic. So, to conclude the models overall for the High School group were weak and not able to distinguish pass from fail.

If we look at the Community College Table, we see that the overall kappa statistic for each model did not go over 0.2. Although the highest correctly classified SMO Base Case at 58.83721% but the kappa statistic was 0.18774 which means the model was weak. As we saw in the High School Group, correctly classified doesn't mean a great model and overall, the CC group models were weak.

A reason that the models may have been weak is because the data sets were too small. Having too little data makes it hard for the algorithm to distinguish what makes an instance a pass or fail. Another reason small data sets may make a weak model is because they don't accurately represent the group at large. The sample may be missing key instances that would help the algorithm distinguish between the two classes of pass and fail.

To add, another reason why the models may have been weak is because of the use of redundant attributes when training the model. For instance, the attributes RaceEthnicity, EthnicCode, CitizenCode, and CountryCCode were used for training HS Group models for the first round of attribute selection. These attributes are redundant because they don't add new information to the training set, this added information is essentially meaningless and is known as noise. Noise makes it harder for the algorithm to discern what attributes make a student more likely to fail.

Its important to note that since the models were not accurate in predicting student outcomes, there no real evidence for any attribute being key markers for success or failure of the WPST.

For future work it is essential to acquire data that is most recent and to train the models on a larger data set. With more data the computer may be able to distinguish between a student who will fail or pass the WPST. It would also benefit to get more information on the students themselves. For example, individual class grades such as English course work may make the notable difference in creating a reliable model. It is also important that further work in attribute selection is necessary. It is imperative that redundancy and noise are mitigated to ensure a better use of the current machine learning algorithms.

Acknowledgements

This project is funded by The Ronald E. McNair Postbaccalaureate Achievement Program which is federally funded through the U.S. Department of Education. A special thanks to CSU-LSAMP which is funded through the National Science Foundation (NSF) under grant #HRD-1826490 and the Chancellor's Office of the California State University. And a special thanks to my Faculty Mentor, Dr. Melanie Martin.

References

Bennett, C. T., Chen, X., & Cataldi, E. F. (2018, February). First-Generation Students College Access, Persistence, and Postbachelor'sOutcomes. Retrieved March 10, 2019, from https://nces.ed.gov/pubs2018/2018421.pdf

Brownlee, J., PhD. (2018, May 23). A Gentle Introduction to k-fold Cross-Validation. Retrieved May 10, 2019, from https://machinelearningmastery.com/k-fold-cross-validation/

Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Fernandes, Holanda,Victorino, Borges, Carvalho, & Erven. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research,94*, 335-343.

Gray, C. C., & Perkins, D. (2019). Utilizing early engagement and machine learning to predict student outcomes. *Computers & Education*, *131*, 22–32. https://doi-org.ezproxy.losrios.edu/10.1016/j.compedu.2018.12.006

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. (2015). A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, 1909-1918.

McHugh, M. (2012). Interrater reliability: The kappa statistic. Biochemia Medica,22(3), 276-282.

Natek, & Zwilling. (2014). Student data mining solution–knowledge management system related to higher education institutions. Expert Systems With Applications,41(14), 6400-6407.

Shahiri, Husain, & Rashid. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science,72(C), 414-422.

Slim, A., Heileman, G., Kozlick, J., & Abdallah, C. (2014). Predicting student success based on prior performance. 2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), 410-415.

Stan State Enrollment at a Glance. (2018, August). Retrieved March 11, 2019, from http://www.csustan.edu/institutional-research/institutional-data/enrollment

The NCES Fast Facts Tool provides quick answers to many education questions (National Center for Education Statistics). (2018). Retrieved March 14, from https://nces.ed.gov/fastfacts/display.asp?id=569

Van der Merwe, C.A., & Van der Merwe, S. (2009). Student success: Data mining measures what matters. 7(2), 275-302..

Vu, Ngan H.(2016). Predictive Modeling of Human Placement Decisions in an English Writing Placement Test, ProQuest Dissertations and Theses.

Writing Proficiency Screening Test (WPST). (n.d.). Retrieved March 3, 2019, from http://www.csustan.edu/wpst

Xu-Ying Liu, Jianxin Wu, & Zhi-Hua Zhou. (2009). Exploratory Undersampling for Class-Imbalance Learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(2), 539-550.