Exploring Individual Feature Importance in Student Persistence Prediction

Zaiyong Tang
Salem State University

Lisa Chen
Salem State University

Anurag Jain
Salem State University

Student persistence is of great importance for all stakeholders in higher education. There have been numerous studies using data mining and machine learning tools to predict student persistence. However, very little research has explored individual feature importance and their distinctive roles in predicting individual outcomes. In this study, we compare the predictive performance of two widely used machine learning models, logistic regression, and random forest, and use the SMOTE to improve the model performance. We analyze the feature importance in both aggregated form and their varied impact on individual predictions using a real-world student persistence dataset. In the discussion section, we propose practical approaches for monitoring and predicting student persistence.

Keywords: student persistence, dropout, prediction, feature importance, individual feature contribution

INTRODUCTION

Colleges and universities around the world have always been interested in student persistence in attaining their education goals. Research studies investigating persistence compose one of the most widely reported areas of research in higher education. The study of student retention can be traced back to the 1600s. It has evolved in modern times from preventing dropouts in the 1960s to building theories in the 1970s, managing enrollment in the 1980s, and broadening horizons with integrated models in the 1990s (Aljohani, 2016). During the last two decades, higher education institutions have faced increased scrutiny and more challenging environmental factors. Online education has gained wider acceptance, especially in the last few years due to the Covid-19 pandemic (Harper and Robison, 2022). However, studies have found that online courses are even more challenging for student persistence (Bawa, 2016).

Recent development in data science and analytics have empowered researchers to build robust predictive models to study student persistence. There are many readily available and easily deployable machine learning software libraries, such as those provided by scikit-learn (https://scikit-learn.org/stable/). Through predictive modeling, we will be able to better understand why some students persist while others fail. (Alyahyan and Düştegör, 2020). There are many factors that impact student persistence, including

student intention, commitment, academic achievement, academic history, high school experience, social integration, and institutional policies (Snyder and Dillow, 2015).

Non-persistence has significant social, economic, and even psychological impacts on individual students as well as on institutions. Identifying students at risk of dropout at an early stage is critical to use intervention measures to help students in need. Assistance can be offered for those at risk of falling behind in multiple ways, such as academic advising, tutoring, mentoring, psychological support, faculty and/or peer interaction, campus engagement activities, etc. The key is to pinpoint the exact areas where individual students need help.

In this research, we use an existing dataset to compare two popular machine algorithms for student persistence prediction. First, we review relevant studies of student persistence. Next, we describe the data and then build a logistic regression model and random forest classification and compare their performance measured by prediction accuracy, in-class recall, and F1-score. We then explore the importance of Individual features. Furthermore, we drill down the prediction for individual students and identify the significant factors that contribute to student dropout prediction. This is followed by discussions of insight we gained in the current study, important issues in modeling student persistence, and further research directions.

Our research is aimed at building an analytical model that can predict student success with high predictive power and identifying individual features that are significant in contributing to successful prediction. Furthermore, we are interested in exploring how those significant features differ when applied to individual student prediction. This information helps to develop targeted and individualized measures to help students at risk. Throughout the paper, we use the terms feature and factor interchangeably.

LITERATURE REVIEW

Continued interest in academic persistence with the widespread acceptance of the Tinto model (Tinto, 1975) has continued unabated into current times. There have been numerous studies and reviews of various student retention and persistence literature in recent years as well. A comprehensive review of student retention models in undergraduate education over the past eight decades is presented by Manyanga et al. (2017). Alyahyan and Düştegör (2020) provide a comprehensive literature review of predicting academic success in higher education. They summarized best practices in predicting student success, from data source identification, data collection, and preparation to data mining tools and analysis. Their goal was to provide a step-by-step guide for researchers and practitioners interested in applying data mining techniques to predict student success.

Sekeroglu et al. (2021) present a systematic literature review of student performance prediction studies between 2010 and 2020. They identified 297 relevant articles from three citation databases. After removing duplicates and publications not meeting the inclusion criteria, a total of 176 articles are summarized. It is interesting to note that in most of the studies, 83.5 percent came in the second half of the decade (2015-2020). This indicates the increased interest in student success research in recent years. The studies were summarized according to their aims, predictive models, datasets, evaluation metrics, and validation strategies.

Rastrollo-Guerrero et al. (2020) carry out a qualitative research study of 64 recent articles on predicting student success. Those articles are summarized based on the objectives and techniques (including methods/algorithms) used. The major objectives are studying student dropout and student academic performance. Only two articles aimed at recommending activities and resources. The major techniques include supervised learning, unsupervised learning, recommender systems (collaborative filtering), artificial neural networks, and data mining techniques.

Chavarriaga et al. (2014) propose a recommender system for students based on social knowledge and assessment data of competencies. The system offers learning advice to students based on an analysis of the student's current competence level against similar former students' performance. Karalar et al. (2021) use an ensemble model for predicting students at risk of academic failure. Their model is based on an ensemble meta-model that combines the prediction by several popular machine learning algorithms, including

gradient boosting, quadratic discriminant analysis, decision tree, random forest, extra trees, logistic regression, and artificial neural network. The advantage of the meta-model is to integrate the strength of individual algorithms that approach the task from different perspectives.

Bawa (2016) reviews the literature of student persistence in online education. They attempt to identify critical factors for high attrition rates in online classes and explore potential solutions to improve retention rates. Although the focus is on online learning, the findings are also pertinent to traditional face-to-face education. The key factors they identified include misconceptions relating to cognitive load, social and family, motivational, technological constraints and the digital natives, lack of instructor understanding of online students, faculty limitations of using technology, and institution limitations to training faculty.

Martins et al. (2021) compare several machine learning models for predicting students' academic success. They tested logistic regression, support vector machine, decision tree, and random forest classifier and found that the random forest classifier outperformed the other model when prediction accuracy and average F1-score were used as the metrics. They note that a common problem in student success prediction is class imbalance. Typically, the dropout/failure rate is significantly lower than the success rate. The authors show that deploying the synthetic minority over-sampling technique (SMOTE) improves the predictive performance of the models.

Batool et al. (2022) provide a comprehensive review of student performance prediction based on approximately 260 studies in the last 20 years. They find that artificial neural networks and random forest classifiers are the most used data mining tools. They also note that feature selection is used before model building by nearly half of the studies. Feature selection is used to remove irrelevant or redundant features so that 1) the prediction results are improved and 2) the model processing time is reduced.

Yağcı (2022) uses random forests, nearest neighbor, support vector machines, logistic regression, naïve Bayes, and k-nearest neighbor algorithms to predict students' final exam grades. The data set consists of records of 1854 students, and the classification accuracy achieved is in the range of 70–75%. The article provides a comparative analysis of 11 recent papers on student success modeling, including modeling objectives, variables, student level, dataset size, algorithms used, and performance results.

DATA DESCRIPTION

The data set used for this study is obtained from ResearchGate. It can be downloaded via the research entry "Predict students' dropout and academic success" at www.researchgate.net. A subset of the data was used in the publication "Early Prediction of Student's Performance in Higher Education: A Case Study" (Martins, et al., 2021).

The data are anonymized undergraduate student data collected between the academic year 2008-2018 at the Polytechnic Institute of Portalegre, Portugal. There are a total of 4424 records and 37 features (variables). The data set has already been pre-processed, removing all records with outliers and/or missing values. Only the dependent variable, named Target, is categorical. Among the independent variables, 28 are numerical and eight are Boolean.

Target represents the outcome of the college students: Graduate, Enrolled, and Dropout. The case study by Martins et al. (2021) processed the data further to classify the students based on the length of time to Success, Relative Success, and Failure. Since our study focuses on student persistence. We combined the original data with the target value Enrolled and Graduate into Persistent.

The 36 independent variables involve mostly academic and demographic data. There are a few macroeconomic and financial variables. Some variables, such as parents' occupation and qualification, may not be easy to explain as we do not have details of the data coding. For example, "Father's qualification" has values ranging from 1 to 44. This study will not deal with the interpretation of the variables. We will focus on the predictive results of the machine learning models and feature impact on the prediction.

STUDENT SUCCESS PREDICTION

To do individualized student persistence prediction and analysis, we need first to build predictive models that support feature analysis. In this section, we compare two popular classification models: Logistic Regression and Random Forest Classifier.

Model Performance Metrics

The performance metrics are based on the ratios of True Positive (the predicted class is the true class), True Negative (the predicted non-class is the true non-class), False Positive (the predicted class is the true non-class), and False Negative (the predicted non-class is the true class). The following are typical measures for classification performance:

- Precision = (True Positive) / (True Positive + False Positive). Percentage of correct prediction for the target class.
- Accuracy = (True Positive + True Negative) / (Total Sample Size). Accuracy gives overall correct prediction across all classes.
- Recall = (True Positive) / (True Positive + False Negative). Percentage of target class overall predicted target class. In other words, recall is the percentage of the class predicted correctly by the model.
- F1 score = 2*(Recall * Precision) / (Recall + Precision).

The F1 Score is the weighted average of Precision and Recall. It is especially useful when the class sizes are uneven. When the class sizes differ substantially, accuracy as a measure might give a false sense of good performance. For example, the dataset used for this study has 4424 students' records with 3003 persistent students and 1421 dropouts. The persistence ratio is approximately 68%. A model that simply classifies all students as persistent would have an accuracy of 68% even if it fails to predict a single dropout.

Logistic Regression Model

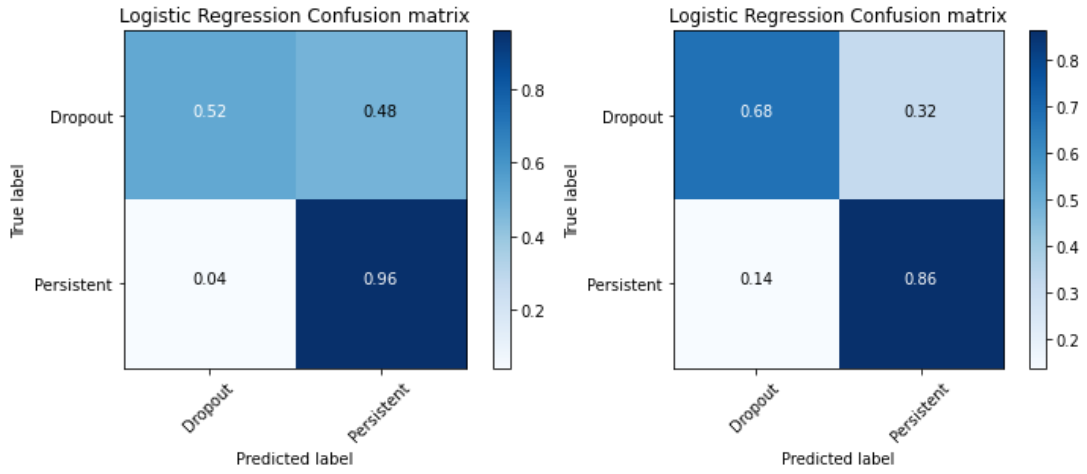
The logistic regression model has been widely used in solving classification problems as it is relatively efficient and easy to implement. We start with logistic regression as a baseline prediction model to predict student persistence. The output of the logistic function lies between zero and one. It can be interpreted as the probability of the data point belonging to the predicted class.

We use the Logistic Regression Classifier from scikit-learn (https://scikit-learn.org/stable/), the free machine learning library for the Python programming language. We randomly split the data into training and testing data sets, with training data consisting of 70% of the total while testing data consisting of 30% of the total data. Training data is used to fit the logistic model, and the model performance is measured with the testing data.

After splitting the data into training and testing data, the training set has a total of 3096 records with 2102 persistent and 994 dropouts. Figure 1 (a) shows the confusion matrix that gives the percentage of correct/wrong classification in the test data. The y-axis represents the actual outcome in the test dataset. The x-axis gives the prediction by the model. Logistical regression produced an impressive 96% correct prediction rate (recall) for the persistent class. However, the recall for the dropout class is only 52%.

As recognized widely by the machine learning research community, data with unbalanced classes may lead to the poor prediction of the minority class. Thammasiri et al. (2014) report that among the three class balancing methods they tested, the synthetic minority over-sampling technique (SMOTE) outperformed random under-sampling and random oversampling. We use SMOTE to balance the training data classes. With SMOTE oversampling of the minority data, synthetic data for the dropout class are created to match the size of the persistent class. Thus, the new training data has a total of 4204 records, with 2102 records for each of the two classes. The test dataset is unchanged. Figure 1 (b) shows the confusion matrix for sampling with SMOTE. The recall increased from 52% to 68% for the minority class, while the recall decreased from 96% to 86% for the majority class.

FIGURE 1
PREDICTING RESULTS OF THE LOGISTIC REGRESSION MODEL

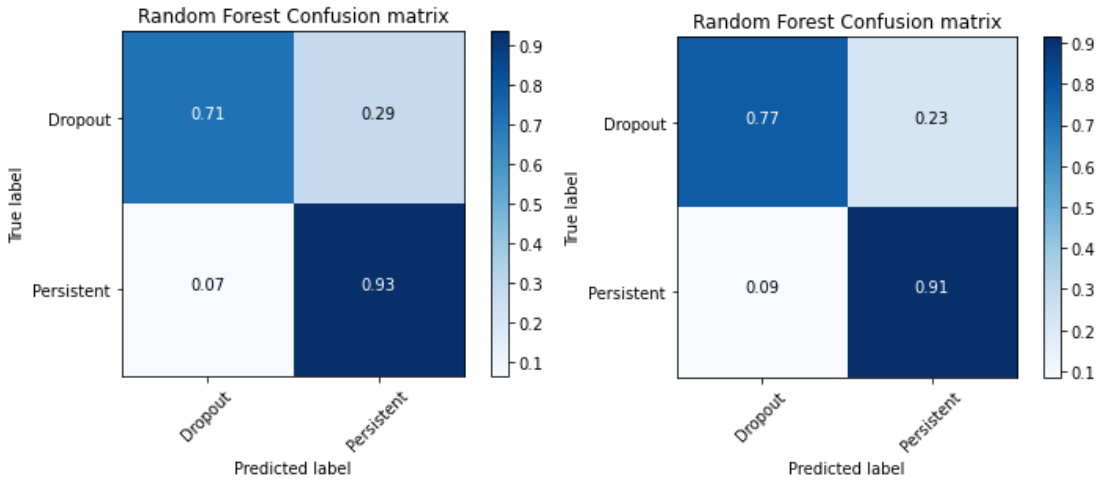


(a) Original Data (b) SMOTE Augmented Data

Random Forest Classifier

Previous studies have noted that the Random Forest model often performs better than the logistic regression model in many cases (e.g., Martins et al., 2021, Yağcı, 2022). We use the Random Forest Classifier from the scikit-learn machine learning library. The same data split, 70% training, and 30% testing are used.

FIGURE 2
PREDICTING RESULTS OF THE RANDOM FOREST MODEL



(a) Original Data (b) SMOTE Augmented Data

Under the same model training and testing conditions used for the Logistic Regression, Figure 2 shows that the Random Forest model outperforms the Logistic Regression model with both the original data and SMOTE augmented data. The recall for the minority class improved to 77%, while the recall for the majority class remained high at 91% when data re-sampling by SMOTE was used.

Model Comparison

The results are based on a random sample split. A different sample split may produce somewhat different results. To compare the two models more accurately, we ran 10-fold cross-validation tests. The data set is randomly divided into equal-sized subsets. For each of the 10 runs, one subset is reserved for testing, while the other nine subsets are used for model fitting.

F1 Score and Accuracy are used as the performance metrics. The results confirm that Random Forest Classifier outperforms the Logistic Regression Model. However, balancing the sample size via SMOTE does not necessarily improve the overall performance measures. It did not improve the F1-score for either model. It did improve the accuracy of the Random Forest model (Table 1)

TABLE 1
MODEL PERFORMANCE COMPARISON VIA 10-FOLD CROSS-VALIDATION

	F1 Score		Accuracy	
	Mean	Std. Dev	Mean	Std. Dev
Logistic Regression	0.882	0.008	0.828	0.012
Logistic Regression (SMOTE)	0.818	0.013	0.807	0.016
Random Forest	**0.912**	0.011	0.877	0.016
Random Forest (SMOTE)	0.905	0.015	**0.906**	0.012

As can be seen in Figures 1 and 2 in the previous section, SMOTE improves the predictability of the minority class. However, the correct prediction of the majority class decreases; hence the overall measures may decrease. For student persistence prediction, correctly predicting the success of the students is valuable. However, correctly predicting the failure of the students is more important as it can lead to impactful actions to alter undesirable outcomes. Hence, in this case, SMOTE deployment is recommended.

INDIVIDUAL FEATURE ANALYSIS

Feature selection plays an important role in building robust predictive models. Choosing the right subset of the features can reduce the undesirable effects of irrelevant variables while preserving model performance (Guyon and Kaelbling, 2003). While most researchers are interested in feature selection before model building, we intend to study the feature importance after building the predictive models with the aim of reducing the number of features. This would make the individualized prediction data analysis, discussed in the following, more amenable.

Feature Comparison of the LR and RF Model

Although the interpretation of the coefficients of the logistic regression is not straightforward, the size and sign of the coefficients indicate the features' impact on the prediction. A large positive coefficient implies a feature's significant contribution to student success, while a large negative coefficient implies a feature's significant contribution to student failure. Figure 3 shows the relative size and sign of the logistic regression model coefficients.

FIGURE 3
LOGISTIC REGRESSION FEATURE COEFFICIENTS

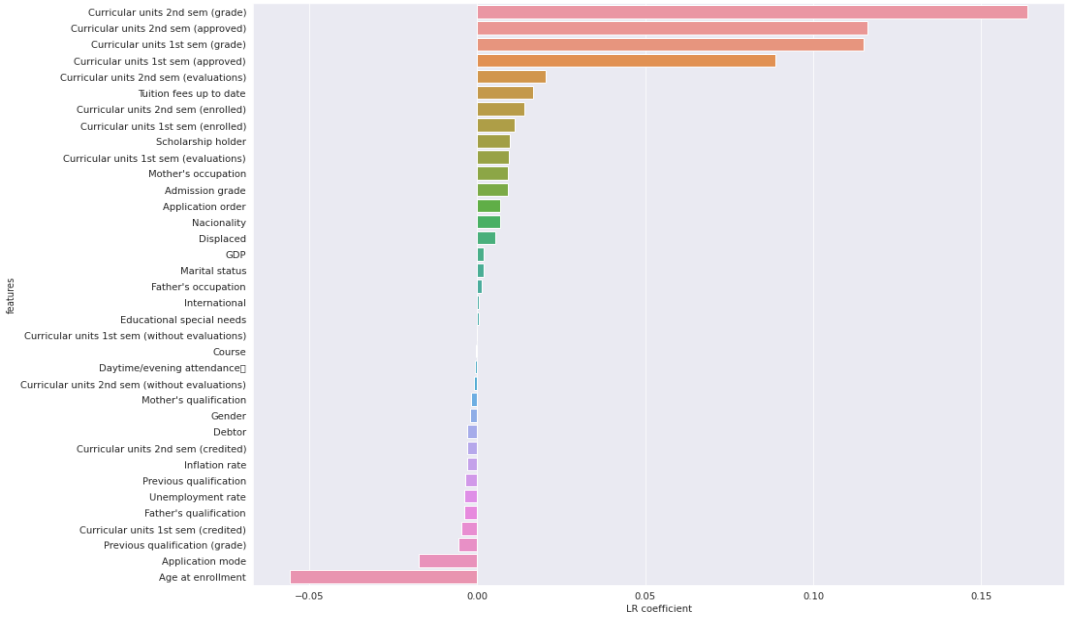


FIGURE 4
RANDOM FOREST MODEL FEATURE IMPORTANCE

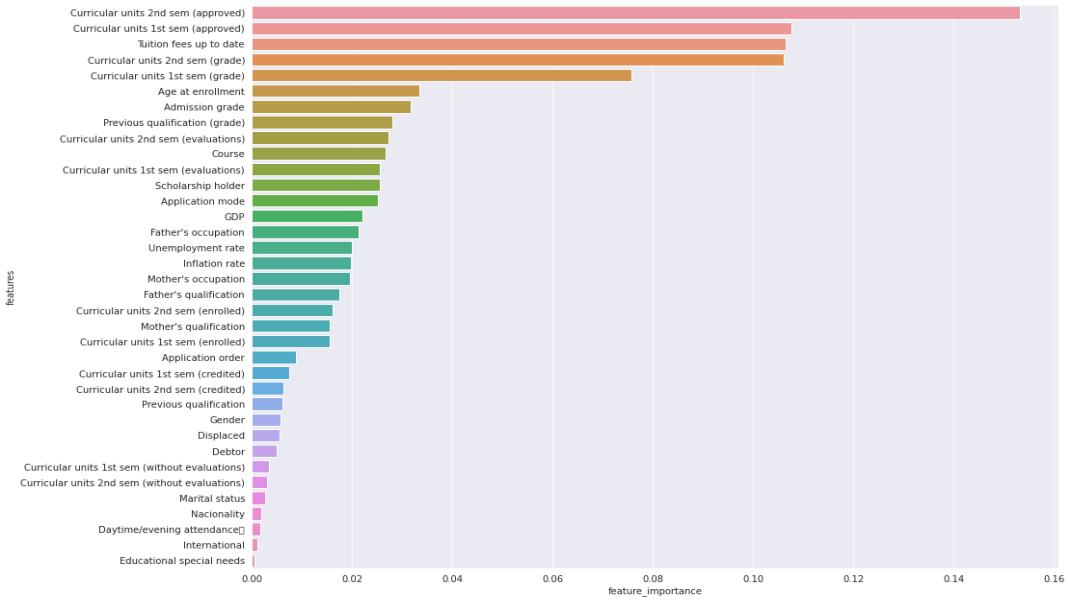


Figure 4 shows the features ordered by their importance for the random forest classifier model. The random forest classifier does not give feature coefficients as the logistic regression model. The model classification is based on synthesizing the decision by many decision trees. However, the algorithm measures the contribution of each feature to the prediction by computing its feature importance.

Comparing Figure 3 and Figure 4, we note that although not all the top-ranked factors are the same by the logistic regression and the random forest model, they are similar. Curricular units of the second and the

first terms are highly ranked. Features with relatively large negative coefficients in the logistic regression are recognized as important by the random forest model as they have a large impact on the prediction outcomes.

Reduced Features

Many of the features are highly correlated, making some of the features unnecessary in building effective predictive models. Since the random forest classifier produces better results than the logistic regression, for the remaining analysis, only the random forest model is used. We tested random forest prediction with a reduced number of features. Starting with the original 36 features, we reduced the feature number to 24, 12, and 6. In Table 2, the results are the averages from 10-fold cross-validation runs using the scikit-learn KFold module. As shown in table 2, the model performance decrease with the reduced number of features is not significant.

TABLE 2
PREDICTION RESULTS WITH DIFFERENT NUMBERS OF FEATURES

Number of Features	F1-Score	Accuracy
36	0.912	0.877
24	0.911	0.874
12	0.905	0.866
6	0.885	0.845

To simplify the analysis of individual student prediction in the next section, we use only the top 12 features (see Table 3) from the 24-feature model. Note that the removed variables are always from the bottom of the feature list ordered by feature importance determined by the random forest algorithm. Also note that the top 12 features of the 36-feature model are not necessarily the same as the top 12 features of the 24-feature model, as the feature importance would be re-computed by the algorithm. Among the 12 features, only "Tuition fee up to date" is a Boolean variable.

TABLE 3
TOP 12 FEATURE PROFILE

Feature Name	Distinct Count	Mean	(Min, Max)
Curricular units 2nd sem (approved)	20	4.44	(0, 20)
Curricular units 2nd sem (grade)	782	10.23	(0, 18.57)
Curricular units 1st sem (approved)	23	4.71	(0, 26)
Tuition fees up to date	2	(binary)	(0, 1)
Curricular units 1st sem (grade)	797	10.64	(0, 18.88)
Admission grade	620	126.98	(95, 190)
Course	17	8856.64	(33, 9991)
Age at enrollment	46	23.27	(17, 70)
Previous qualification (grade)	101	132.61	(95, 190)
Curricular units 2nd sem (evaluations)	30	8.06	(0, 33)
Mother's occupation	32	10.96	(0, 194)
Curricular units 1st sem (evaluations)	35	8.30	(0, 45)

Feature Contribution to Individual Prediction

Most student persistence studies stop at aggregated prediction results. However, no student is the average student. We need to not only predict the likelihood that a particular student will persist but also

identify individual features that contribute to the prediction. Especially when the predicted probability of success is low, then the contributing features would enable us to identify areas where potential intervention actions can be applied.

The feature importance given in Figure 4 is the aggregated importance derived from the entire training data. However, for individual students, each feature may play a different role in predicting their persistence. We deploy the Tree Interpreter (https://pypi.org/project/treeinterpreter/), a software library for interpreting predictions by decision tree and random forest algorithms of scikit-learn. It helps us find the contributions of individual features to a particular individual prediction.

After training the random forest model with the SMOTE augmented data and the 12 selected features, the trained model is used to predict the success (persistence) of all students in the testing dataset. For each prediction, the feature contributions to the prediction are sorted by their importance to student success. Figure 5 shows a case of prediction of 90% probability of success, represented by the blue bar on the right-hand side. Green bars represent a positive contribution to success, while red bars represent a negative contribution. The feature names are shown on the x-axis. The y-axis represents the probability. The height of the bar represents the proportion of the contribution. Note that the base number is 0.5; that is, the probability of success is 50% without any feature contribution. This is because the two outcomes (success and failure) have the same sample size due to the use of SMOTE.

FIGURE 5
FEATURE CONTRIBUTIONS TO A PREDICTION OF SUCCESS

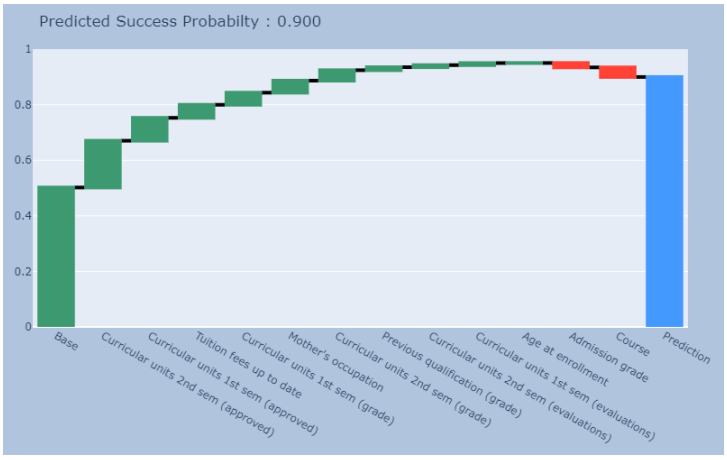


Figure 6 shows the sample of predictions with a low probability of success (towards failure). With a classification threshold of 0.5, those two individuals are predicted to drop out. Note that the feature importance order is different for the two individuals. For (a), the admission grade contributes significantly to potential success. However, tuition and fees not up to date contribute heavily to lowering the probability of success. The next two significant negative factors are second-semester curricular units approved and grade. For (b), the most significant positive factor is the second-semester curricular unit's grade, while the two most significant negative factors are previous qualification and second-semester curricular units approved.

FIGURE 6
FEATURE CONTRIBUTIONS TO THE PREDICTION OF FAILURE

(a) Success probability at 0.18

(b) Success probability at 0.45

As the source of the data is secondhand, we are not able to dive deep into the interpretation of the features. For example, how the semester curricular unit grade, evaluations, and approved are related. However, this approach of drilling down to individual prediction analysis is useful in general. The detailed and precise information about individual cases enables us to develop impactful interventions to help students to reduce the risk of failure.

Top Contributors to Dropout

Using the test data (30% of the total, 1328 records), we focus on the factors that contribute to student dropout. As shown in the previous section, individual students may have different factors that push them toward failure. For the two-class classification (target is zero or one) problem, standard procedures typically use 0.5 as the threshold to decide the predicted outcome. The correct prediction of dropout by the random forest classifier is 76.1% (Table 4). The slight difference between this value and the value in the section on feature analysis is due to the reduced number of features used in the prediction. The random tree initialization of the algorithm may also cause the prediction outcome to change slightly.

For intervention measures that aim at improving student persistence, the correct prediction of student failure is more significant than the correct prediction of student success, as the consequences of a wrong failure prediction could be more serious. If we increase the classification threshold, we can improve the dropout prediction significantly.

TABLE 4
CORRECT PREDICTION RATE AT DIFFERENT THRESHOLD

Prediction Threshold	Correct Prediction Rate (%)	
	Dropout	Persistent
0.50	76.11	90.11
0.60	80.56	85.13
0.70	84.78	77.47
0.75	87.35	71.59

Using a 0.6 classification threshold, we tallied the ranking of contributing factors for all the correctly predicted dropouts. Of the 427 dropouts in the test data, 344 are predicted correctly. As shown earlier, multiple factors (not always the same) contribute to individual predicted results. However, it is also important to identify which features are affecting most students. The most significant factor contributing to the predicted failure is "Curricular units 2nd sem (approved)," approximately 69 percent of the predicted dropouts. The number two feature that is the most significant for predicted failure is "Tuition fees up to date," approximately 22 percent of the predicted dropouts.

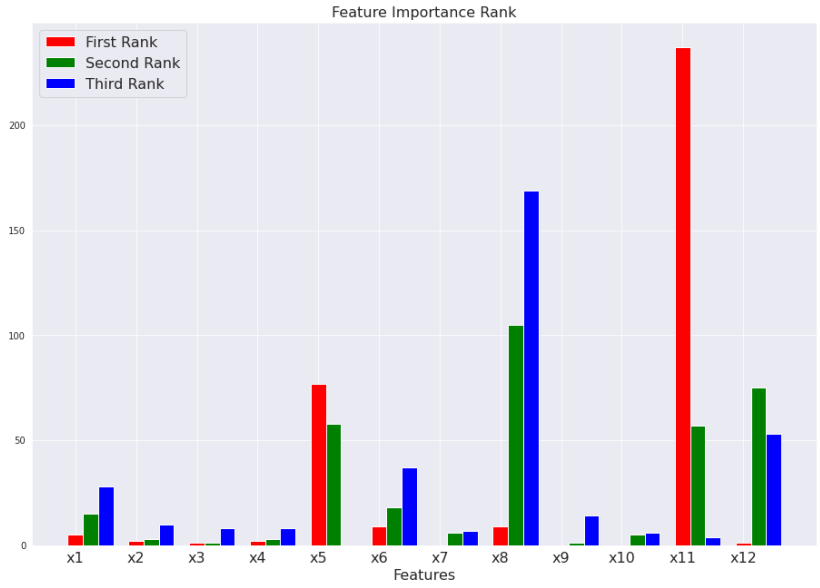
Table 5 shows the list of features and the counts of the number of times they are the first, second, and third largest contributors to the predicted dropout. The first, second, and third contributors correspond to the largest, the second largest, and the third largest red bar in the individual feature contribution charts in Figure 5 and Figure 6. The order of the features in the table is based on the order of the variable names in the original dataset, with variables renumbered from x1 to x12.

TABLE 5
FEATURE RANK (FREQUENCY) IN PREDICTING DROP OUT

Variable	Feature Name	First	Second	Third
X1	Course	5	15	28
X2	Previous qualification (grade)	2	3	10
X3	Mother's occupation	1	1	8
X4	Admission grade	2	3	8
X5	Tuition fees up to date	77	58	0
X6	Age at enrollment	9	18	37
X7	Curricular units 1st sem (evaluations)	0	6	7
X8	Curricular units 1st sem (approved)	9	105	169
X9	Curricular units 1st sem (grade)	0	1	14
X10	Curricular units 2nd sem (evaluations)	0	5	6
X11	Curricular units 2nd sem (approved)	237	57	4
X12	Curricular units 2nd sem (grade)	1	75	53

To get a better visual of the case numbers in Table 5, we put the feature importance ranking in a bar chart in Figure 7. X11 is the leading contributor to dropout, while X8 appears most frequently as the second and third leading contributor to student failure. Again, we do not have the data coding details to dive into a deep analysis of those variables and explain why they play such an important role in dropout prediction. However, the approach we use here can give us more insightful information when data coding details become available.

FIGURE 7
FEATURE IMPORTANCE FOR DROPOUT PREDICTION



DISCUSSION

Significant efforts have been put in place to improve student retention and persistence by many colleges and universities. Studies have shown that it is far more cost-effective to retain current students than successfully recruit new students to replace the lost ones (McGinity, 1989). One of the key components of student risk management and prevention is to be able to monitor student progress in real-time. Thus, predicting student persistence based on historical data is not enough. We need data analytics and predictive models that provide real-time results.

There are tools available to analyze student data, show historical trends, and project future outcomes. For example, Navigate from EAB (formerly Education Advisory Board). Navigate is an enterprise-level student success management system that integrates with student information systems to provide real-time data analytics about individual students. EAB claims that Navigate is used by more than 850 colleges and universities, and it serves more than 10 million students. However, EAB's predictive model is proprietary. In other words, it is a black box that provides student risk estimates but does not provide information on how such an estimate is derived. From Navigate, we know if a student is at risk of failure, but we do not know why. Thus, the value of such information is limited.

Academic performance is a leading indicator of student persistence. However, there are many other factors, such as social-economic, demographic, community engagement, activities, advising, peer support, etc. We propose building a comprehensive predictive model that (1) incorporates a wide range of features, (2) provides real-time dynamic data analytics, and (3) drills down to individual feature contributions to student success prediction, as demonstrated in the previous section. With this type of detailed and individualized information, we would be able to customize student support services so that student retention efforts become more effective.

Integrating student information systems with data visualization tools such as Tableau (https://www.tableau.com/) and Power BI (https://powerbi.microsoft.com/en-us/) can provide real-time feedback on student progress and pinpoint areas that might cause concerns for student success. Again, we emphasize the importance of drilling down to individual student levels of data and predictions. This type of tool can be tailored to serve the needs of multiple clienteles, from college administration to faculty, staff, and students.

Student data come from various domains such as academic, demographic, financial, and activities/engagement. Although all relevant student data should be collected, cleaned, stored, and analyzed for proactive prevention strategies and actions, the focus will be on the academic and engagement variables, as there is not much we can do about demographic variables other than recognizing they may play a role in student success. Obtaining the right data take enormous effort. Some data, such as student activities and community engagement data, may not be captured by the institutions. Or they are captured and stored in different systems that are separated from traditional student (academic) information systems that are under different authorities of the organizational units. Thus, gathering and maintaining all relevant student data can be very challenging.

The current research is based on a limited student dataset. For further research, we intend to consider data with more diverse features, particularly those data linked to potential risk intervention measures. We would also consider developing prototype data visualization that incorporates the individualized predictive model. Such a system can provide real-time information about a student compared with peer groups and identify areas that flash red light for potential problems. Another way to develop a better predictive model is to perform rigorous feature analysis and selection before model building. This is especially important when we collect a wide range of student data with many features.

CONCLUSION

We carried out an extensive analysis of a student persistence data set, from aggregated prediction to feature importance to individual prediction data and feature analysis. The aim of this research is to build powerful predictive models that give us actionable insights so that we can drill down to individual students and find out where they stand and how we can help them. The main contributions of this paper are (1) comparing the predictive performance of logistic regression vs. random forest classifier with class-biased student data and unbiased SMOTE augmented data; (2) identifying the feature importance in the aggregated model, and (3) analyzing feature contributions in individual student dropout predictions. In the discussion section, we proposed practical approaches for monitoring and predicting student persistence.

REFERENCES

Aljohani, O. (2016). A Comprehensive Review of the Major Studies and Theoretical Models of Student Retention in Higher Education. *Higher Education Studies*, *6*(2), 1–18.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*(1), 1–21. https://doi.org/10.1186/s41239-020-0177-7

Batool, S., Rashid, J., Nisar, M.W., Kim, J., Kwon, H.-Y., & Hussain, A. (2022). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies.* https://doi-org.corvette.salemstate.edu/10.1007/s10639-022-11152-y

Bawa, P. (2016). Retention in Online Courses: Exploring Issues and Solutions–A Literature Review. *SAGE Open*, *6*(1).

Chavarriaga, O., Florian-Gaviria, B., & Solarte, O. (2014). *A Recommender System for Students Based on Social Knowledge and Assessment Data of Competences*. Springer International Publishing. https://doi-org.corvette.salemstate.edu/10.1007/978-3-319-11200-8_5

Guyon, I., Elisseeff, A., & Kaelbling, L.P. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(7–8), 1157–1182. https://doi-org.corvette.salemstate.edu/10.1162/153244303322753616

Harper, J.C., & Robinson, J.B. (2022). Teaching from a Distance: Challenges in Classroom Management to Promote Professionalism. *Journal of Business & Educational Leadership*, *12*(1), 35–56.

Karalar, H., Kapucu, C., & Gürüler, H. (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International Journal of*

Educational Technology in Higher Education, *18*(1), 1–18. https://doi.org/10.1186/s41239-021-00300-y

Manyanga, F., Sithole, A., & Hanson, S.M. (2017). Comparison of Student Retention Models in Undergraduate Education from the Past Eight Decades. *Journal of Applied Learning in Higher Education, 7,* 30–42.

Martins, M.V., Tolledo, D., Machado, J., Baptista, L.M.T., & Realinho, V. (2021). *Early Prediction of student's Performance in Higher Education: A Case Study*. Springer International Publishing. https://doi.org/10.1007/978-3-030-72657-7_16

McGinity, D.A. (1989). *A Path of Analysis of Effects of Multiple Progress of Student Persistence*. Ann Arbor, MI: University of Michigan.

Rastrollo-Guerrero, J.L., Gomez-Pulido, J.A., & Duran-Dominguez, A. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. *Applied Sciences-Basel*, *10*(3), 1042. https://doi-org.corvette.salemstate.edu/10.3390/app10031042

Sekeroglu, B., Abiyev, R., Ilhan, A., Arslan, M., & Idoko, J.B. (2021). Systematic Literature Review on Machine Learning and Student Performance Prediction: Critical Gaps and Possible Remedies. *Applied Science-Basel*, *11*(22), 10907. https://doi-org.corvette.salemstate.edu/10.3390/app112210907

Snyder, T.D., & Dillow, S.A. (2015). National Center for Education Statistics (ED), & American Institutes for Research. (2015). *Digest of Education Statistics 2013. NCES 2015-011*. National Center for Education Statistics.

Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014, January 1). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, *41*(2), 321–330.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*, 89–125

Yagci, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi-org.corvette.salemstate.edu/10.1186/s40561-022-00192-z