ESTIMATION OF PERSISTENCE AT A COMMUNITY COLLEGE: A COMPARISON OF ALTERNATIVE MACHINE LEARNING MODELS

Fermin Ornelas

Institutional Research, Rio Salado College, Maricopa Community Colleges, Tempe, AZ, USA.

ABSTRACT

This research focuses on developing persistence models for Rio Salado College. It is an initial effort to predict persistence from one term to the next. Several ensemble models are experimented and compared in their respective key metrics such as: confusion matrix, AUC, F1-Score, and feature importance. Exploratory data analysis is undertaken to narrow the set of variables utilized in the models. Two models were considered for possible implementation: a logistic regression and a gradient boosting machine. The former is easier to implement and explain to non-technical personnel, while the latter behaves like a black box. Based on key performance metrics, the model of choice was the gradient boosting machine. Development and testing were conducted with python using jupyter notebooks. The author hopes that this experimental process will fill a vacuum in the analytical needs of community colleges.

KEYWORDS

accuracy, AUC, confusion matrix, ensemble models, feature importance, gradient boosting machines, Machine Learning, persistence, precision, recall

1. INTRODUCTION

In the current economic conditions, academic institutions continue making efforts to adapt to the student needs. The pandemic surge followed by high levels of unemployment has made remote working and learning necessary alternatives for both workers [2], [3] and students to continue providing for their families and enhancing their working and academic skills, respectively.

In previous recessions, higher education institutions experienced countercyclical enrolment increases, but during the current COV19 driven recession that has not been the case. Higher education is currently having to reduce personnel or put on hold additional hiring and is suffering from enrolment declines due to the pandemic and degrading economic conditions, [15]. According to [22], overall enrolment is down by 3% with freshman enrolment being down to 13%. These declines are more severe for 2-year colleges than for 4-year institutions. Community colleges according to [23] data are experiencing a 19% decline with a pronounced decline among students of colour. Moreover, a recent survey conducted by an education marketing company [14], among 528 students across the U.S. concluded that 43% of prospective students for one and 2-year programs are considering delaying enrolment. Other findings of the survey were: 54% of traditional age students considered having a degree or certificate extremely valuable; 65% of non-traditional students because of the pandemic thought that a degree or certificate was extremely valuable. Considering these facts, by necessity educational institutions are increasingly creative in trying to retain and support their enrolled students. The prolonged pandemic and the sudden

recession have caught everyone by surprise and some not well prepared for working remotely while providing training and education to a smaller student population in an online environment.

After the Great Recession, all institutions, experienced dynamic student growth reaching a peak around 2010. Over 8 million students enrolled at 2-year colleges [16]. Since then, by 2017, enrolment at these institutions has declined by about 1 million [5]. It is this scenario that has all institutions and Rio in particular, focusing their efforts on retaining current students. Henceforth, Rio Salado College has been innovative in finding ways to leverage predictive analytics to enhance academic success and persistence among its diverse student population. It has developed a course success tracking system, Rio PACE, to provide timely feedback to students and instructors. In addition to its course success predictive modelling strategies [24], [27], the researcher is currently working on another modelling project on persistence. The objectives are to understand student's behaviour and leverage empirical results to foster persistence and retention while leveraging Rio Salado College resources to effectively support its student population.

Persistence, a dichotomous event, for this project is defined as enrolling in a term (Fall 2021) and continuing into the next term (Spring 2022). However, [23] defines persistence as continuing studying at any institution for a second year. In addition, the report concludes that 6 in 10 students persist into the second fall term. The definition to be used here obeys two reasons: a concern for students' continuity in their studies and the fact that the majority of Rio College student population are not pursuing an associate degree or certificate. That has created two types of students: those attending to take specific courses often to transfer the credits to a major university; and a two-prong segment pursuing 1-year certificates or associate degrees. Thus, this investigation will seek to identify features that motivate non-degree, degree, and certificate seekers to persist or not into the next term.

There is also a technical complexity associated with development, implementation and tracking of machine learning models. Anxious in their quest to serve students better, colleges are contracting with vendors to effectively attend to their student population needs with timely advice and to prevent further enrolment declines. Those vendors, often suggest complex solution packages that 2-year colleges cannot afford or may not fit their analytical needs. That has resulted in some discomfort on the use of such applications and subsequent recommendations [6], [13], [29]. These solutions behave like black boxes understood only by vendors' technical personnel. In a recent paper [4], it is reported that researchers and college administrators have little practical means of assessing predictive software results potentially having adverse effects for students. Furthermore, we intend to compare analytical solutions for various ensemble classifiers: logistic regression, decision trees, gradient boosting, random forest, ADABOOST, and XGBOOST. The research will be developed using easily available software tools and LMS inhouse data, hoping to contribute to the analytics needs of 2-year college institutions.

Therefore, the current investigation project has the following objectives: to develop persistence a model for two populations, degree and certificate, and non-degree seeking students; to identify features that drive student persistence; and to compare results from all the different models. Therefore, the article will discuss literature review, followed by methodological aspects of the model, next it will present empirical findings, followed by limitations of the study. Finally, it will render conclusions and recommendation arising from this research.

2. LITERATURE REVIEW

Predicting persistence at community colleges has always been a challenging issue. Some studies have attempted to predict persistence using pre-enrolment high school variables and focus on first time in college students. Most have attempted to predict persistence focusing on a sense of belonging and involvement at 4-year institutions, [18], [30]. This research focuses on persistence research and estimation issues related to community colleges using machine learning modelling.

In [4] a study to evaluate whether students at the Virginia Community College system graduated with a college level credential within six years of initial entry. They compared alternative models on two dimensions: random truncation of a current cohort sample aligned to enrolment length distribution of historical cohorts, and different variable construction, i.e., term specific and constructed. The pursued models were, OLS regression, Cox proportional hazard, random forest, and XGBoost. Over 331 predicting features fitted into the models utilized to predict graduation. [4] concluded the models predicted reasonably well. However, the predictions differed among models in the students' probability ranking, depending on the alternative variables tried. They expressed concerns that this variability could result in inefficient use of resources for targeting students and potential bias against underserved students. [4] observed that no substantial increases in accuracy existed when applying more complex models. In another interesting research [28], as part of their modelling task to predict MOOCs dropouts' online behaviour, built predictive features based on their expertise and crowd sourcing. Some in the latter group were provided an artificial data set and encouraged to write their own scripts for feature engineering. This approach resulted in a richer set of features.

The self-proposed approach rendered 18 features, while crowd sourcing resulted in another 10 new covariates. Some of these variables had high predicting power. These features were tested on students enrolled in a Circuit and Electronics course during Fall 2012. Randomized logit was the selected model and students were divided into four cohorts: passive collaborator, wiki contributor, forum contributor, and fully collaborators. Predictive power of the engineered features varied by cohort.

In [25], an investigation to test a comprehensive list of factors impacting student retention at the institutional level. Employing a logit and a random forest model they discovered that GPA, institution's primary campus, first generation in college, and academic advisor were the primary factors impacting student retention behaviour. The random forest model had, in descending order, primary campus location, first generation, age, GPA, and academic advisor as the top five features important predictors of student retention. Overall model accuracy was reported at 78%, no other metric was provided. The logit model revealed that academic advisor, support services, primary campus location, and academic achievements were relevant for student retention.

Focusing on 9,200 first time community college students (FTIC) over a four-year period, [8], studied fall-to-spring and fall-to-fall retention. The time window started from Fall 2001 to Fall 2004. They found that developmental education and online courses had impact on student persistence. Other predictors were, financial aid, parent's education, ethnicity, hours enrolled and dropped per term in the first fall semester, and support services participation. Stronger predictors for fall-to-spring retention, in descending order, were: passing developmental reading, taking an online course, participating in student services, not taking a developmental reading course, passing a developmental math course, receiving financial aid, father with some college, hours enrolled in the first fall term, and student age. Unfavourable odds of retention were: not taking developmental math, mother's education, and hours dropped in the first fall term. Similarly, for fall-to-fall retention, about the same variables were found to positively impact retention, except

for passing a developmental writing course and mother with college education. Age was found to be non-significant in predicting retention.

In another early study, [7], identified predictors of attrition before community college students began classes by focusing on pre-enrolment factors. Using a logistic classifier, main findings were: high school GPA was the strongest predictor, the lower the GPA the greater the odds of dropping out, age indicator 20-24, being minority except for Asian and enrolled part-time. Interestingly, the researcher found no impact for remedial course in reading, math, and writing. Tracking revealed that students planning to graduate, or transfer had higher retention rates. A caveat on models with ethnicity variables, minimizing bias would require further refinement when predicting non-retention of minorities [4], [6].

In a retention study at CSU - San Bernardino, [15] focused on yearly retention from freshmen to sophomore year for first time freshmen entering in the fall 2009 and fall 2010 quarters. They wanted to identify predictors of retention and the likelihood of students dropping out. The final logistic classifier included the following features: ethnicity, high school GPA, university studies 100 enrolment, first term GPA, percent of courses completed in the first year, and number of GENED course taken in the first year. Decile probability scores were bucketed into high, medium, and low risk of not being retained. At-risk students would be encouraged by peer advisors to engage in university activities such as: service and community learning, diversity/global experiences, learning communities and undergraduate research.

At NYIT, [2], report on an end-to-end model implementation to predict first-year retention. The data collected for this process came from admission applications, registration/placement testing, surveys on students when taking the Compass placement test, and financial information on 1,453 students during the fall of 2011 and 2012. Based on this sample a year later only 983 returned, 68% retention rate. The models tried were logistic regression, neural networks, naïve Bayesian, decision tree, and ensemble. The latter was the model of choice with recall values of 73% and precision of 54% in the validation data set. Furthermore, the prediction results were shared with counselling staff who would use the report for at-risk student treatment.

To improve retention and academic achievement at a community college, [17], conducted a treatment and control experiment where students in the former group attended a freshmen seminar taught by specially trained instructors/advisors while the control group were taught by conventional instructors. Results proved that those in the treatment group achieved higher GPA and were retained into the immediate term at a higher rate than those in the control group. The experiment was applied to 280 students, 158 and 122 were in the treatment and control group, respectively. The 14 sections for the course were limited to traditional sections only during the Fall 2009 and Spring 2010.

In another study at a community college in Southeast Texas, [12] evaluated the impact of a student success course (SSC) on three metrics: persistence, retention, and academic achievement. Employing the CCSSE survey for items 4 and 9 in a sample of 432 students, 197, and 235 in the treatment and control groups, respectively. [12] conducted Chi-square testing for independence concluding that students enrolled in the SCC program had higher persistence, and retention rates, and better academic achievement in gatekeeping courses. No attempt was made to build a predictive model, however that does not diminish the relevance of the study nor the impact of the SSC on the three outcome measures.

At the University of South Florida, [19], utilizing pre-matriculation characteristics and data derived from the College Student Expectations Questionnaire built a logistic classifier for

retention under two alternatives: one based on the former type of data; and another one using both pre-matriculation and survey data. The sample population was incoming FTIC, 3998 students in the Fall and Summer of 2006. The retention rate for Fall 2007 was reported at 82.2% for the model version with pre-matriculation data and 78.2% for the model using both pre-matriculation and survey data. Correctly predicted observations were reported as 82.3% and 78.9% under each alternative model, respectively.
For the first model, variables having a positive effect on persistence were as follows: high school GPA, Asian vs White, Black vs. White, and time elapsed since orientation. Adverse role was reported for SAT combined, pre-nursing major, and residence status commuter. For the second model, the factors positively predicting persistence were, high school GPA, being black vs. being white, expecting to participate in clubs and student organizations, expecting to read textbooks or assigned books in college, expecting to work on campus. Negatively influencing persistence were, expecting to read non-assigned books, and expecting to work off-campus while in college. [19], followed this work with some suggestions for creating interventions to treat at risk of not-persisting students.

In another persistence study at community colleges [21] intended to explore likely factors impacting students' decision whether to drop or stay in school. The authors found that cumulative GPA was the strongest predictor of persistence followed by units taken and English proficiency. Other attributes such as age, work hours, and financial aid were initially statistically significant, but their performance diluted once other variables were entered into the predictive model. This phenomenon was also observed for psychosocial and academic integration variables.

3. ENSEMBLE MODELS AND METRICS

The models developed for this research belong to supervised learning. The main characteristic of these classifiers is the existence of a target outcome, in our case, persistence. While logistic regression is based on the sigmoid function, the other models are derived from decision trees. The latter are referred as nonlinear nonparametric classifiers because they do not have any distributional assumptions. They can either tune their learning algorithm or the dataset used to train them, to ensure diversity and high model performance [11].

Simple models are more likely to introduce bias as they are unable to fit the data because of underfitting, while complex models are prone to introduce more variance by overfitting the data. Ensemble models are intended to address both bias and variance by combining predictions of multiple base models. Models such as bagging, ADABOOST, random forest, gradient boosting, and XGBOOST employ weak learning and majority voting to arrive at a final model selection often using decision trees as base models. They are considered generative because they can generate and affect the base learners they use, [11]. Weak learning is a sequential process where the errors of the weak predictions are assigned a higher weight compared to correctly classified predictions when estimating the base models. The errors from these weak learners are used in subsequent estimation. The objective is to improve the overall predictions. Majority voting or average voting requires comparison of results from each base classifier to make a final selection, based on a majority agreement on the outcome of interest prediction. In addition to enhance the final classifier results often a researcher applies hyperparameter tuning. These are employed either using grid search or random search techniques to fine tune the final model.

The machine learning model results discussed in this research are compared in their performance metrics. The confusion matrix is often a 2x2 contingency table representing both correct and misclassified predictions. True positive and true negatives are identified along the diagonal, while false positives and false negatives are found in the off-diagonal cells. These values are utilized to

compute the following metrics: recall or true positive rate, specificity or true negative rate, false positive rate, and precision.

A classification report built by Python sklearn metrics application summarizes these metrics by outcome class. Another important metric is the ROC curve, it is a plot of the true positive rate against the false positive rate at different cut-off points for the classifier. Commonly, a threshold of .5 is established to classify a prediction outcome as either success or failure. In our case, a probability value above that threshold would mean a student has persisted, otherwise he/she has not. Note that this value can be changed but it will also affect the prediction classification rates.

There is also another measure that eases model interpretability, feature importance. It gives a percentage ranking of the features selected by the model and can be used to illustrate users the impact that those features have on the outcome score. Therefore, this research will demonstrate the use of these measures to arrive at a final model recommendation. Since model interpretability is a concern among operational personnel, there are new applications intended to address it. New packages on such as LIME or SHAP are intended to look deeper in the prediction results and the modelling features contributing to the predicted probability helping to erase the notion that machine learning models behave like a black box [20]. These techniques will not be utilized in this research.

4. EXPLORATORY DATA ANALYSIS

The sample data for model development and validation was extracted from the inhouse MLS using MSQL queries. The total cleaned sample encompassed 19,459 records for the Fall 2021 and the next term was Spring 2022. Of these, 80% was utilized for model development and 20% remained for validation, 15,567 and 3,892 observations, respectively. The actual persistence rate for this data is reported at 53% and 47% as non-persisting. These figures avoid issues of unbalanced data and ease the estimation process. As part of the exploratory data analysis (EDA), descriptive statistics were calculated for both samples. Common engineered features were dummy transformations for categorical features. A full-time/part-time flag, and a Rio Pace flag were created. Variables derived from previous education and age. The latter was binned in categorical ranges, while the former was used to create bins for different education type groups. There were no missing values in the data, but some of the outliers for the continuous variables were replaced using the first and third quartile whisker values for credits attempted, cumulative earned hours, and cumulative quality hours. Frequencies and bar charts were undertaken to get a better understanding of the data. Those results are not presented in this paper. Distributional plots were also pursued for the continuous attributes. However, the research only presents bar, box, plots, and densities relevant for the modelling process and final selection.

A visualization of the correlation matrix among the features selected for modelling is provided below in figure 1. The correlation chart shows most variables are not strongly correlated among themselves or to the outcome of interest, except for the fulltime flag and credits attempted at .69. The variable cumulative quality hours is also correlated with cumulative earned hours at .57. Negative correlation is present between the Rio PACE indicator and age younger than 22 years, at .52. The rest of the coefficients are below .5.

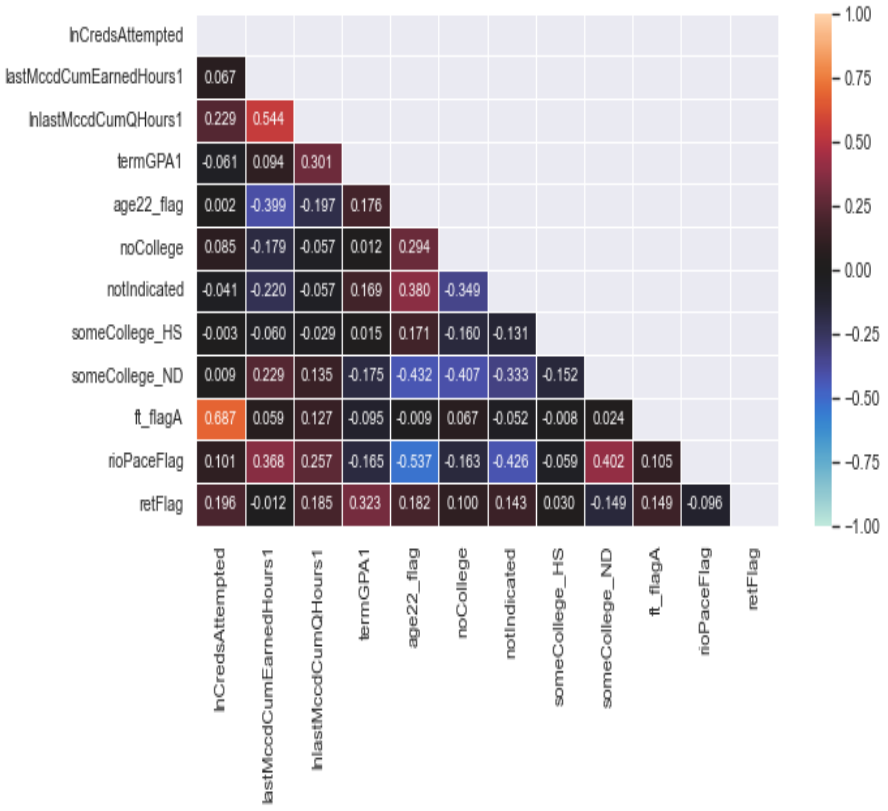

Figure1. Correlation Matrix for Selected Features

Boxplots are also provided for the continuous variables -credits attempted, credits earned, and term GPA without outliers. There appears to be statistical differences among the variables between persisting and non-persisting students in those three variables. In all instances, persisting students take and earn more credits. Subsequently, they manage to achieve a higher term GPA. The median values for credits attempted and earned and term GPA are larger for persisting than for non-persisting students. Because the continuous variables appear to be non-normally distributed, statistical testing was conducted using the Wilcoxon non-parametric test for those variables at alpha=.05. The tests concluded that there were statistical differences in the means of persisting and non-persisting students. Not shown in the analysis were the Chi-square testing for the categorical variables with respect to the persistence indicator. All variables were statistically significant suggesting that their inclusion in the modelling effort was appropriate. Also, collinear diagnostics were conducted using the variance proportion factors and the condition index using SAS.

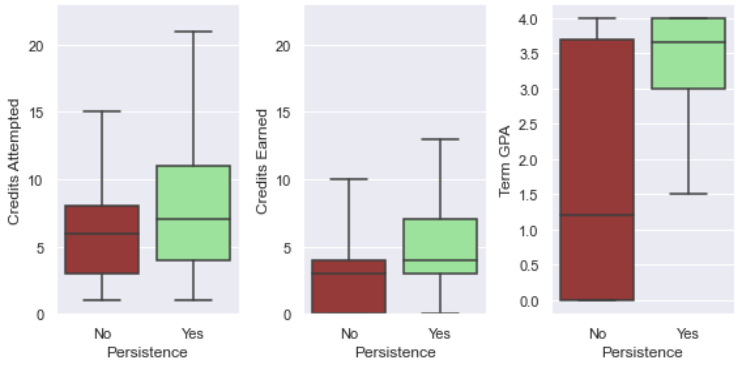

Figure 2. Boxplots for Credits and Term GPA

These results are relevant to instructors and advisers because they signal student at risk of not persisting. It is worth mentioning that as part of the exploratory analysis numerical variables included in the model were treated for outliers and credits attempted was logarithmically transformed.

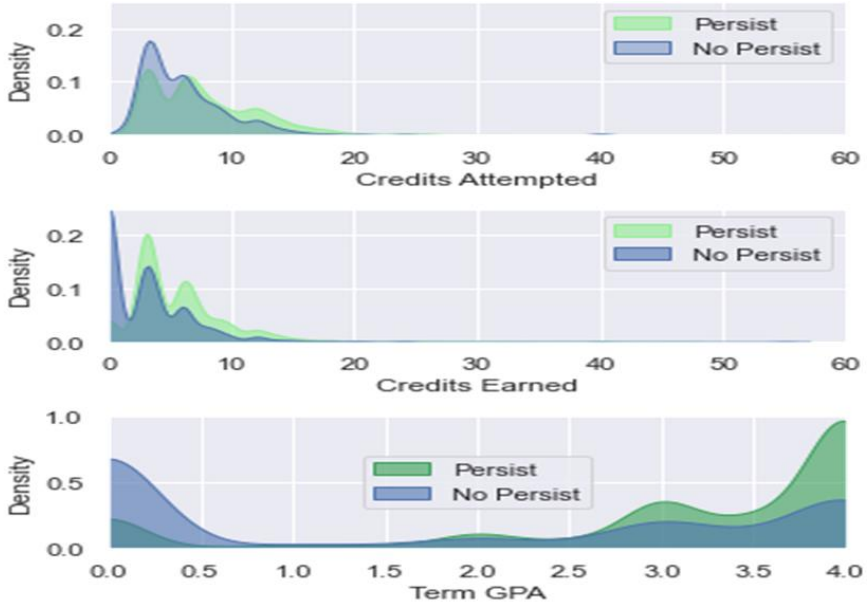

Figure 3. Credits and Term GPA

Density plots are shown in figure 3 above, the credits attempted chart suggests that non-persisting attempt about 3 credits more often than persisting students, but after about 6 credits non-persisting students lag persisting ones. The second chart shows that persisting students earn more credits than non-persisting students also. The skewed tail exists because some students are enrolled in law enforcement training and take more courses than regular degree and non-degree seeking students. For term GPA, the left tail of the distribution shows that non-persisting students have problems maintaining a passing GPA, while the right tail demonstrated that persisting students outperform non-persisting ones at around 2.5 GPA or higher. Thus, the chart strongly suggests that term GPA could influence the student decision on whether to continue his/her studies into the next term. This fact is likely to be reflected in the model result and in some of the literature reviewed.

These descriptive findings will help strengthen the case for building a machine learning model for persistence. The following section presents model results, all models estimated are visually compared on key performance metrics: confusion metrics, feature importance, AUC, accuracy, precision, recall, and F1-score. Based on the results, emphasis is given to results for the gradient boosting machine (GBM) model.

5. EMPIRICAL RESULTS

To analyse persistence into the next term at Rio Salado college, we undertook various alternative models: logistic regression, decision trees, bagging, ADABOOST, random forest, gradient boosting machine, and XGBOOST. Ten-fold cross-validation comparison results suggested that decision trees and bagging were the worst performers. The rest of the models are compared in their performance metrics such as feature importance, area under the curve, accuracy, precision, recall, and F1-score. A final model will be selected based on these metrics for possible implementation and to provide prediction probability scores for current students.

Table 1. Models Comparison Results

(a) Training of Models

Metric	Logistic Regression	Gradient Boost Tuned	XGBOOST Tuned	ADABOOST Tuned	Random Forest Tuned
Accuracy	0.7369	0.8627	0.6723	0.6980	0.8557
Recall	0.7891	0.8753	0.9944	0.8726	0.8659
Precision	0.7354	0.8676	0.6196	0.6646	0.8631
F1	0.7613	0.8714	0.7635	0.7545	0.8645

(b) Validation of Models

Metric	Logistic Regression	Gradient Boost Tuned	XGBOOST Tuned	ADABOOST Tuned	Random Forest Tuned
Accuracy	0.7433	0.7903	0.6477	0.7002	0.7824
Recall	0.7860	0.8029	0.9787	0.8739	0.7932
Precision	0.7453	0.8029	0.6042	0.6663	0.7967
F1	0.7651	0.8029	0.7472	0.7561	0.7950

Tables (a) and (b) provide the set of performance metrics to be compared for each alternative model for training and validation. Sample sizes were: 15567 and 3892, respectively. Often for decision tree-based models, estimates for the training data set tend to overfit the data. In these results only, the random forest model appears to suffer from slightly overfitting. For this reason, training results are not very useful in making a model selection. It is necessary to evaluate validation results as they tend to simulate environment production conditions. Training results in table (b) suggest that both random forest tuned and gradient boost tuned have similar performance with a slight advantage for the gradient boost tuned model.

The gradient boost tuned managed to achieve better performance metrics than all the models. Recall, precision, and F1-score are about 80%, while accuracy is at 79%.

Table (a) and figure (4) demonstrate that random forest and gradient boosting tuned have similar performance metrics; behind these models are the results for the logistic regression classifier. Interestingly, XGBOOST achieved the highest recall but its performance decays in the rest of the metrics. The gradient boosting model metrics remain in the range of 80%, except for accuracy at about 79%. The random forest follows with slightly lower metrics in the range of 79% with a similar value for recall. Logistic regression performance is slightly lower in the 74- 78% range. Its recall value is at 78%. XGBOOST and ADABOOST reached the highest recall rate, 98 and 87%, respectively, but their performance in other metrics were lower.

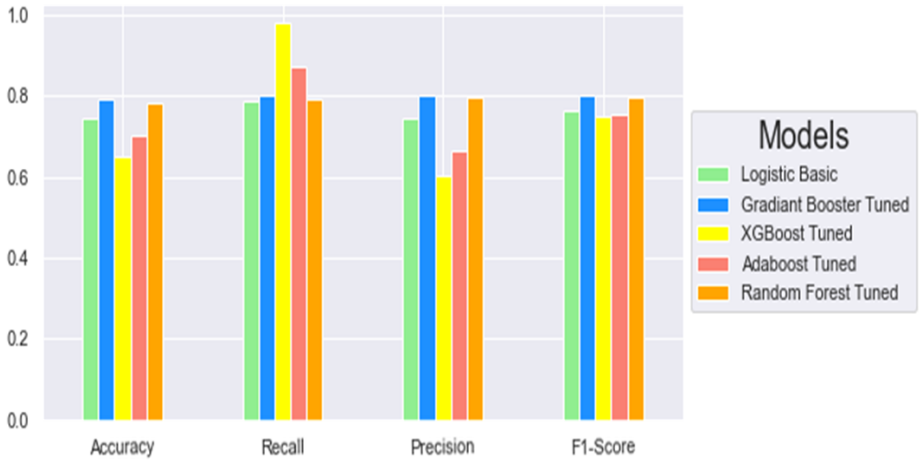

Figure 4. Model Metrics Performance on Validation Data

For those reasons, the researcher is inclined to suggest three models to test for possible implementation: gradient booster, random forest, and logistic regression. Next, we present the AUC curves for these tree models to show why the gradient boot machine is the preferred model of choice. Validation results for the three models are overlaid in the same chart. The farther to the left the better the model classifies the validation data. The 45-degree line represents business as usual, a 50-50 chance that someone persists or not.

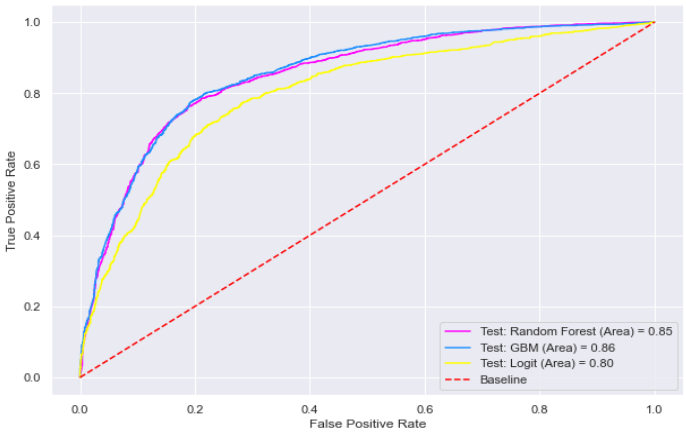

Figure 5. ROC Curve Validation Results Comparison

In figure5 above two models have very close AUC performance. Gradient boost machine and the random forest in their tuned version achieved 86 and 85% values, accordingly. The logistic model while also having a decent performance coefficient, it trails both models with an AUC value of 80%. Ideally, the closer the curves are to the upper left corner, the better the model is in its classification of the outcome.

Next, we present both the confusion matrix, figure5 below, and the feature importance chart, figure6, for the gradient boosting machine model. The true label represents the actual cases in the validation dataset, while the predicted values are the predicted assignment of those cases. The latter, gives a good approximation of model performance in a production-like environment. By assessing the misclassification values in the confusion matrix, one is led to conclude that this model does discriminate better between persisting and non-persisting students.

The main diagonal figure demonstrates the correct classification for non-persisting and persisting students. One can observe the model does have a higher classification of persisting students compared to non-persisting ones, i.e., 42.7% versus 36.3%.

Regarding the off-diagonal values instructors and advisors could be interested in the false positive rates as this figure would represent cases that are not persisting identified as persisting. Doing so, could lead to non-persistence prevention if accompanied by an intervention targeting student at risk of not persisting. On the other hand, false negative rates create a false alarm and could cause misallocation of scarce resources since these students are likely to persist with minimal or no intervention.

There is no harm in the misclassification of persisting students as non-persisting, however there will be resource misallocation if targeted for interventions such as intrusive advising or additional student support. On the contrary, identifying non-persisting student as persisting could be a missing opportunity, especially at a time when 2-year colleges are experiencing enrolment declines.

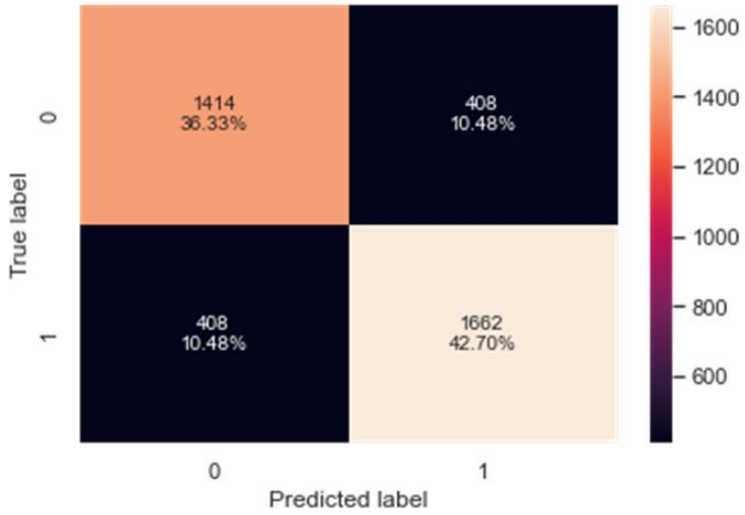

Figure6. GBM, Confusion Matrix

Finally, we think it is important to illustrate what factors contribute to persistence. To address this issue, the feature importance chart is presented below.

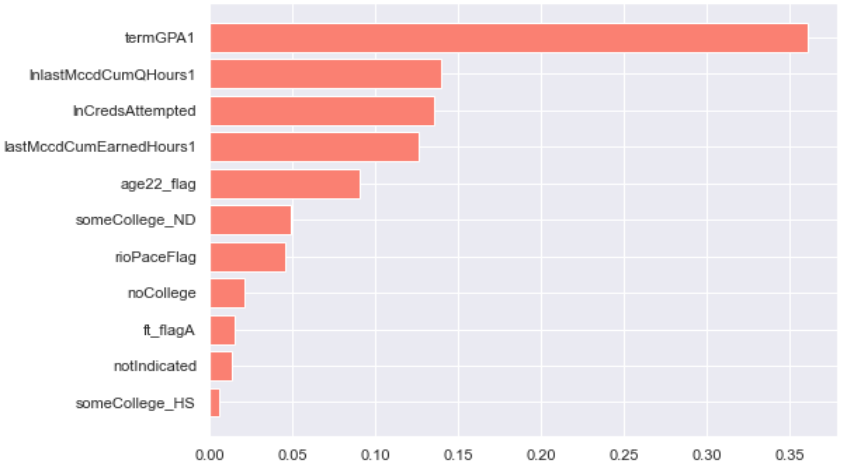

Figure 7. GBM Feature Importance

Ranked by importance one can observe that term GPA is the most important attribute that influences whether a student will enrol into the next term, its impact is over 40%. Next, log transformed credits attempted is the second most important feature with a coefficient on around 14%. Quality cumulative hours, cumulative earned hours, and age indicator of 22 years or less have around 10% contribution to the probability of the outcome of interest. After that the rest of the features have importance coefficients at 5% or lower.

Overall, one can observe that the variables belong to three groups: academic performance, education background, and age. In the next section, we extend the support for possible model implementation to the gradient boosting machine model because it presents slightly better results on misclassification rates.

6. LIMITATIONS OF THE STUDY

The study focused on estimating student persistence into the next term. It explored alternative machine learning models and compared their key metrics. The data was split into training and validation. However, having a third data set for testing would have made the results more robust as it would have eliminated the possibility for data leakage [10]. The latter is a potential problem arising from trying to improve the model to enhance validation results. Moreover, the number of variables used in the modelling effort could be expanded by engineering additional features as suggested by [4] and [28], that is something to be explored in the future. Efforts will be made to implement and track model results in a production environment. Despite these limitations, the model selected has higher metric values than some of the models cited in the literature review, [2], [25].

There are other studies comparing machine learning models with different attributes selected. [26], attempted to identify factors responsible for hearth failure. The outcome of interest indicator had three levels: mild, moderate, and severe. The algorithms utilized were: CART, NN, and SVM. In the end, the CART model outperformed the other models with an accuracy of 84%. The comparison is similar but the models tested are not the same. Furthermore, this research is among the few ones using machine learning models applied to community college data using readily available open-source software such as Python and Jupyter notebooks as IDE.

7. CONCLUSIONS AND RECOMMENDATIONS

A data set was extracted for students who took courses in the Fall of 2021 and Spring 2022. The purpose was to assess persistence into the next term. Seven classifiers were initially explored: logistic regression, Bagging, Random Forest, GBM, ADABOOST, XGBOOST and a decision tree. In the end, only five models appeared to be worth of additional exploration: logistic regression, Gradient Boosting, random forest, ADABOOST and XGBOOST.

Model performance was compared across the five models on key performance metrics for the training and validation datasets: accuracy, precision, recall, and F1-Score. In the training dataset, the random forest classifier appeared to achieve higher metrics. The gradient boosting classifier was second with lower metrics performance, while the logistic regression classifier performance was third. For the validation dataset, the gradient boosting model secured higher values in all the key metrics. This was followed by the forest random forest and logistic regression classifiers. Therefore, results suggest that the gradient boosting model would be the model of choice for possible implementation.

Of the models evaluated we selected two for consideration: the gradient boost (GBM) and the logistic regression. The former performs the best of all models, but the latter is easier to interpret and implement. In both instances term GPA appears as the most important feature in explaining student persistence. Transformed credits attempted is the second most important variable in the GBM model, while in the logit has the college not indicated as the second most relevant. The third most relevant is the transformed cumulative quality hours in the GBM, while in the logit it is no previous college experience. Confusion matrix results, performance metrics, AUC chart, and variable importance led the researcher to recommend Gradient Boosting Machine as the model choice.

REFERENCES

[1] Aceujo, Esteban M., Frech J., Ugalde, Araya, M. P., & Zafar B. (2020). The impact of COVID-19 on student experience and expectations: Evidence from a survey. Journal of Public Economics 191.

[2] Agnihotri, L. & A. Ott. (2014). Building A Student At-Risk Model: An End-to-End Perspective. Proceedings of the 7th International Conference on Educational Data Mining.

[3] Altig, Dave, Baker S., Barrero, J. M., Bloom, N., Bunn, P., Chen S., Davis, S. J., Leather, J., Meyer, B., Mihaylov, E., Mizen, P., Parker, N., Renault, T., Smietanka, P., & Thwaits, G. (2020). Economic Uncertainty Before and During the COID-19 Pandemic. Journal of Public Economics 191.

[4] Bird, A. Kelli, Castleman, L. B., Mabel, Z., & Song, Y. (2021). Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education. (EdWorking Paper: 21-438). nages

[5] Community college enrollment crisis? Historical Trends in Community College Enrollment. AACC, 2019.

[6] Ekowo, M., Palmer, I. (2016). The Promises and Peril of Predictive Analytics in Higher Education. A Landscape Analysis. New America, Oct. 2016 Fain, P. Top of the Mountain? https://www.insidehighered.com/news/2011/12/21/community-college-enrollment-growth-ends.

[7] Feldman, M. J. (1993) Factors Associated with One-Year Retention in a Community College. Research in Higher Education, Vol. 34, No 4, 1993.

[8] Fike, D. S. & Fike, R. (2008). Predictors of First-Year Student Retention in the Community College. Volume 36, Number 2. October 2008, 68-88.

[9] Juszkiewicz, J. (2020, July). *Trends in Community College Enrollment and Completion Data, Issue 6*. Washington, DC: American Association of Community Colleges.

[10] Kapoor, S. & Narayanan, A. (2022). Leakage and the Reproducibility Crisis in ML-based Science. arXiv:2207.07048v1 [cs.LG] 14 Jul 2022

[11] Kyriakides, G. & Margaritis, K. G. Hands-on Ensemble Learning with Python. Packt. www.packt.com.

[12] Kimbark, K., Peters, M. L., Richardson, T. (2016). Effectiveness of the Student Success Course on Persistence, Retention, Academic Achievement, and Student Engagement. Community College Journal of Research and Practice. http://dx.doi.org/10.1080/10668926.2016.1166352

[13] Klempin, S., Grant, M., Ramos, M. (2018). Practitioner Perspectives on the Use of Predictive Analytics in Targeted Advising for College Students. CCRC Working Paper No. 103. May, 2018.

[14] Lane Terralever. The Pandemic's Impact on Higher Education Marketing in 2020 and Beyond. https//www.laneterralever.com/industries/higher-education-marketing-agency/higher-education-marketing-white-paper-pandemic-impact.

[15] Lopez-Wagner, M. C., Carollo, T., Shindledecker. Predictors of Retention: Identification of Students At-Risk and Implementation of Continued Intervention Strategies.

[16] Inside Higher Ed. December, 2020. Higher Ed Faces Steep Cuts with Recent Oil Bust. https://www.insidehighered.com/news/2020/12/16/higher-ed-faces-steep-cuts-recent-oil-bust

[17] Ryan, M. G., Improving Retention and Academic Achievement for First-Time Students at a Two-Year College. Community College Journal of Research and Practice, 37: 131-134, 2013.

[18] Miller, J. E. & Berger, J. B. A Modified Model of College Student Persistence: Exploring the Relationship Between Astin's Theory of Involvement and Tinto's Theory of Student Departure. Journal of College Student Development; Jul/Aug 1997; 38, 4; ProQuest Education Journals pg. 387.

[19] Miller, T. E. and Herreid, C. Analysis of Variables to Predict First-Year Persistence Using Logistic Regression Analysis at the University of South Florida, 2008.

[20] Mishra, P. (2022) Practical Explainable AI Using Python. https://doi.org/10.1007/978-1-4842-7158-2.

[21] Nakajima, M., A., Dembo, M. H., Mossler, R. Student Persistence in Community Colleges. Community College Journal of Research and Practice, 36: 591-613, 2012.

[22] N. S. C. Stay Informed with the Latest Enrollment Information. November, 2020. https://nscresearchcenter.org/stay-informed/

[23] N. S. C. The Role of Community Colleges in Postsecondary Success: Community Colleges Outcomes Report. nscresearchcenter.org.

[24] Ornelas, F. and Ordonez, C. (2017). Predicting Student Success: A Naïve Bayesian Application to Community College Data. Tech Know Learn 22:299-315. DOI 10.1007/s10758-017-9334-z.

[25] Parvez, R. & Chowdhury, N. H. K. (2020). Economics of Student Retention Behavior in Higher Education. Paper presented at the Agricultural and Applied Economics Association Meetings, 2020.

[26] Prudvi, P. S., Sharifahmadian, E. Applying Machine Learning Techniques to Find Important Attributes for Heart Failure Severity Assessment. International Journal of Computer Science Engineering Applications (IJCSEA) Vol 7, No 5, October 2017.

[27] Smith, V.S., Lange A., & Huston, D. R. (2012). Predictive Modeling to Forecast Student Outcomes and Effective Interventions in Online Community College Courses. *Journal of Asynchronous Learning Networks*. Vol. 16, Issue 3.

[28] Taylor, C., Veeramachaneni, K., O'Reilly, U. Likely to Stop? Predicting Stopout in Massive Open Online Courses. arXiv:1408.3382v1 [cs.CY] 14 Aug. 2014.

[29] The Institute for College Access & Success (2019). Don't Stop Improving: Supporting Data-Driven Continuous Improvement in College Student Outcomes, March 2019.

[30] Tinto, V. (1988). Stages of Student Departure: Reflections on the Longitudinal Character of Student Living. Journal of Higher Education, Jul. – Aug., 1988, Vol. 59, No. 4 (Jul. – Aug., 1988), pp. 438-455.

AUTHORS

FERMIN ORNELAS holds a Ph.D. in Agricultural Economics from Texas A&M University, College Station TX, USA, and a Certificate in Data Science and Business Analytics from the McCombs School of Business at the University of Texas in Austin TX, USA.

He is currently a Sr. Research Analyst at Rio Salado College. His professional research interests include: machine learning modelling applied to student behaviour and the financial industry, i.e. student success, persistence, and retention, credit risk, and customer loyalty.