BOARD # 279: NSF IUSE: Handling Imbalanced Engineering Persistence Data in Machine Learning with Undersampling & SMOTE

Arinan De Piemonte Dourado, University of Louisville

Arinan Dourado, Ph.D., is an Assistant Professor of Mechanical Engineering at the University of Louisville. Prior to joining UofL, he worked as a Lecturer in his home country (Brazil) for three years, teaching and mentoring low-income, first-generation STEM students from rural communities. Additionally, Dr. Dourado worked as an instructor at the University of Central Florida for two years, primarily serving Hispanic first-generation students. Currently, his working on developing and applying machine learning/artificial intelligence tools to identify and suggest intervention actions to increase student retention and success.

Christian Zuniga-Navarrete, University of Louisville
Alvin Tran, University of Louisville
Luis Segura, University of Louisville
Dr. Xiaomei Wang, University of Louisville

Xiaomei Wang is an Assistant Professor in the Industrial Engineering Department of University of Louisville. She received her PhD in Industrial Engineering from University at Buffalo.

Dr. Campbell R Bego, University of Louisville

Campbell Rightmyer Bego, PhD, PE, studies learning and persistence in undergraduate engineering programs in the Department of Engineering Fundamentals at the University of Louisville's Speed School of Engineering. She obtained a BS from Columbia University in Mechanical Engineering, a PE license in Mechanical Engineering from the state of New York, and an MS and PhD in Cognitive Science from the University of Louisville. Her current interests are generative AI and information literacy for engineering, individualized, first-year persistence interventions, and the effectiveness of evidence-based practices in the engineering classroom.

NSF IUSE: Handling Imbalanced Engineering Persistence Data in Machine Learning with Undersampling & SMOTE*

Arinan Dourado[1], Christian Zuniga-Navarrete[2], Alvin Tran[3], Luis Javier Segura[2], Xiaomei Wang[2], and Campbell Bego[4]

[1]Mechanical Engineering, University of Louisville
[2]Industrial Engineering, University of Louisville
[3]Computer Science and Engineering, University of Louisville
[4]Engineering Fundamentals, University of Louisville

Abstract

This work-in-progress focuses on the completed Phase 1 of a funded NSF-IUSE project employing explainable machine learning (ML) models to predict engineering attrition while identifying malleable factors for individualized targeted intervention. Over the course of three years, three common ML models (namely neural networks, random forest, and deep Bayesian networks) will be trained using 5 years of retrospective data from a large southeastern university, in conjunction with global and local ML explanations methods, to identify students at risk of attrition and determine the main malleable factors that lead to such predictions. This funded project essentially aims at developing a general ML framework capable of answering which students are at risk and why, so that targeted interventions can be proposed to meet individual needs. The completed phase includes data preprocessing and preliminary predictive models testing. The findings addressed in this paper are related to the issue of data imbalance in ML modeling. Datasets related to engineering persistence often exhibit significant imbalance,i.e., far fewer students leave programs than persist. This imbalance poses challenges for ML model training and evaluation, often skewing predictions toward the majority class (students who persisted). This short paper and poster presents how our project addressed these challenges by employing a combination of random undersampling and Synthetic Minority Over-sampling TechniquE (SMOTE). The combined effect of these techniques improved the performance of the ML classifiers considered.

Introduction

Student attrition in engineering programs poses significant challenges to educators and policymakers. Research has revealed that students decide to leave engineering for many interrelated reasons (e.g., [1–4]), making it challenging to intervene and help many students at the same time. Iden-

*This project is funded by NSF IUSE: EDU, Award 2335725

tifying at-risk students and understanding the factors contributing to their decisions is critical for developing effective interventions. Machine learning (ML) techniques provide powerful tools for predictive analytics, offering the ability to detect patterns and insights from complex datasets. This funded effort focuses on three ML models: neural networks, random forests, and Bayesian models, which are applied to identify engineering students at risk of attrition. Additionally, explainable ML methods, specifically Local Interpretable Model-Agnostic Explanations (LIME) [5] and SHapley Additive exPlanations (SHAP) [6], are later employed to interpret these predictions. To ensure robust analysis, techniques addressing data imbalance—a common issue in engineering persistence datasets—were evaluated in Phase 1 and will be the focus of this paper.

Neural networks (NN) [7] are inspired by the structure of the human brain and consist of layers of interconnected nodes (neurons) that process input data to generate outputs (e.g., classification outputs persistence/attrition). These networks learn by adjusting the weights of connections through iterative training using labeled examples. Neural networks are particularly adept at capturing non-linear relationships in data. For example, in predicting student attrition, neural networks can identify intricate interactions between factors such as grades, demographic variables, and psychological factors. Random forests (RF) [8] are an ensemble learning method based on decision trees. A decision tree predicts outcomes by following a series of binary splits in the data based on specific criteria (e.g., test scores above or below a threshold). Random forests combine multiple decision trees to create a robust model, where each tree contributes to the final prediction through majority voting (for classification) or averaging (for regression). Given that RFs are an aggregation of multiple decision trees, they usually are less sensitive to data noise and/or incomplete data when compared to other ML models. Bayesian models, particularly deep Bayesian networks (DBN), bring a probabilistic perspective to ML. DBNs extend traditional neural networks by incorporating uncertainty estimates into their predictions, and when incorporated with a generative stochastic model (such as Bernoulli Restricted Boltzmann Machines, see [9]) that learns a probability distribution over inputs, becomes especially useful in capturing latent structures in data, making them valuable for analyzing student persistence patterns.

In engineering persistence datasets, data imbalance often arises because the majority of students persist. This imbalance can bias ML models toward the majority class, reducing their effectiveness in identifying at-risk students. To mitigate this, techniques such as random undersampling [10] and Synthetic Minority Oversampling Technique (SMOTE) [11] can be used. *Random Undersampling* reduces the size of the majority class by randomly removing samples, balancing the dataset. While simple, it risks discarding potentially valuable information from the majority class. *SMOTE* generates synthetic examples for the minority class by interpolating between existing minority samples. This approach preserves the majority class's diversity while augmenting the minority class, leading to more balanced training data. By employing these strategies, ML models can achieve better performance and fairness, ensuring that predictions are not disproportionately influenced by the majority class.

Methods

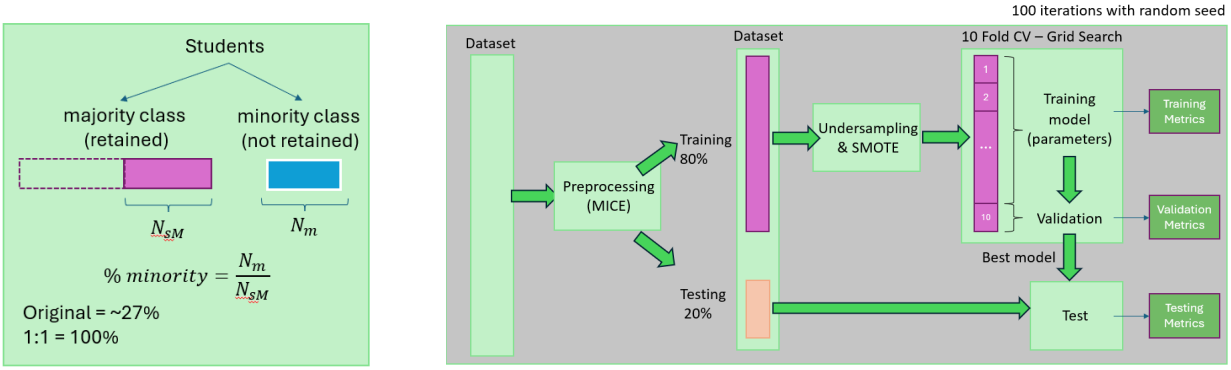
Participants: Our study utilized retrospective data collected from the years 2018-2022. In total, N = 2440 engineering students responded to the surveys. A preliminary investigation showed that around 30% of our students (N = appx. 723) left by the second year. Persistence here is defined

as enrollment in the engineering school in the fall of the second year. The data in this preliminary study included the following variables:

- Demographic Data: gender, race, Pell Grant eligibility
- Survey Data (collected at the beginning of the freshman year Fall semester and again at the end of the same semester): individual interest, perceived effort, opportunity, and psychological costs, perceived academic competence, self-efficacy (surveys can be made available upon request)
- Performance Data: ACT scores (composite, English, math, science reading), term 1 engineering course grades (math, introduction to engineering, and chemistry)
- Financial Aid: source (federal, state, institutional, private), type (scholarship, loan, grant, work-study), and cause (need, merit)

The data was split into two-time stamps: early and later data. Early data contained information available within the first month of the semester: early survey responses, demographic information, financial aid, and ACT scores. Later data included all of the early data in addition to end-of-semester survey responses and student performance in their courses.

Procedures: All variables were preprocessed before being fed into the ML models. Numerical attributes were standardized (i.e., mean 0 and standard deviation 1) and the categorical attributes, such as gender, were converted to numerical variables (e.g., binary). Depending on the model, some categorical features were one-hot encoded to increase performance. Previous research shows large differences for some models based on encoding strategy [12]. Multiple Imputation by Chained Equations (MICE) method, a statistical method for data imputation, was used to handle any missing data (e.g., incomplete survey responses). Then, a split of 80% of the available data was used for ML models training and validation, while the remaining 20% was reserved to test model performance. Varying levels of random undersampling ratio (see Figure 1a) in conjunction with SMOTE were employed to derive the training datasets. Undersampling is first used to create the described class ratios and SMOTE is subsequently employed to balance the yielded dataset. Each training set derived from a given undersampling ratio went to a training loop of 100 iterations (each with a distinct initial condition, i.e., random seed) to improve tuned model parameters robustness. On each training loop k-fold cross-validation (a technique that divides the dataset into k equally sized folds, using $k-1$ folds for training and the remaining fold for validation) was adopted to minimize the chances of model overfitting. Subsequently, the ML models were trained to classify student persistence, where student attrition was labeled as 1 and student persistence was denoted as -1. Figure 1b provides a detailed description of the overall training procedure.
The classification results were assessed using common classification metrics, namely, accuracy (proportion of correct predictions made by the model across all instances, students who persisted and those who did not), recall or sensitivity (ratio of correctly identified at-risk students, emphasizing model's capacity to identify as many at-risk students as possible), precision (ratio reflecting model's ability to minimize false alarms, e.g., 80% precision indicates that 8 out 10 students flagged by the model are actually at-risk), and f1-score (precision and recall harmonic mean).



(a) Adopted undersampling strategy. (b) ML models training and testing procedure.

Figure 1: a) Undersampling percentages allude to the ratio of students who did not persist compared to students who persisted during training. b) Overall model training and testing (evaluation of never-seen data) procedure.

Results and Analysis

Figure 2 illustrates the trend of the NN model performance when the proposed data balancing strategy is employed. The values presented are related to the test set (the portion of available data reserved for model evaluation) and considering later data. The observed spread is due to the multiple iterations during the training loop, with the focus on the ML model's median performance (red bar of the boxplot). While accuracy and precision do not vary much with the proposed approach, a significant increment is observed in recall and f1-score values as also noted in Table 1. Similar trends are observed with early data but with lower performance metrics values.

For the purposes of the funded project, the goal is to achieve high f1-sores, given that this metric is tied to the ML models' ability to correctly identify as many at-risk students as possible while reducing false positives and unnecessary allocation of resources to students who are not genuinely at risk. The obtained results demonstrate that the proposed data balancing strategy had either a positive (RF and NN) or neutral (DBN) impact on the ML models' f1-scores. The greatest gain was observed in the NN model, which went from a median value of 46% in the original dataset to 83% with a $1:1(100\%)$ undersampling ratio + SMOTE (see Tab. 1). The RF model showed similar behavior, with $\sim 20\%$ increment in recall and marginal gains in f1-score, at a detriment of marginal losses in accuracy and precision. Such behavior is expected given that the proposed data balancing procedure impacts the available information concerning the majority class (students who persisted). This means that to increase the models' ability to correctly identify at-risk students, some not-at-risk students will be incorrectly flagged (thus impacting the accuracy and precision metrics). The DBN model presented marginal variations in its performance metrics, suggesting not being sensitive (neutral impact) to the proposed strategy.

Hence, the proposed data balancing strategy either increased the considered ML models' ability to correctly identify students at-risk of attrition or at least did not negatively impact model performance (see the DBN case).

Table 1: Median (and max.) values of performance metrics in the testing set considering later data. Bold cells indicate the best values for each ML model, while * highlights the best overall values.

Model	Original				Undersampling 1:1				SMOTE				Under. + SMOTE			
	Acc.	**Recall**	**Prec.**	**F1**	**Acc.**	**Recall**	**Prec.**	**F1**	**Acc.**	**Recall**	**Prec.**	**F1**	**Acc.**	**Recall**	**Prec.**	**F1**
RF	**82**	58	**75**	65	79	**76**	60	67	80	70	63	67	80	74	62	**68**
	(85)	(65)	**(78)**	(71)	(82)	**(82)**	(64)	(72)	(83)	(80)	(66)	(72)	(83)	(77)	(67)	**(72)**
NN	79	32	84	46	84	79	88	84	83	75	**89***	81	**84***	**78***	88	**83***
	(81)	(44)	(74)	(58)	(85)	(79)	(90)	(84)	(87)	(81)	**(91)**	(86)	**(86)**	**(80)**	(91)	**(85)**
DBN	77	**72**	56	**63**	76	70	55	61	**77**	70	**58**	63	76	67	55	60
	(77)	**(75)**	(60)	**(66)**	(76)	(70)	(55)	(61)	**(77)**	(71)	**(58)**	(64)	(77)	(71)	(55)	(61)

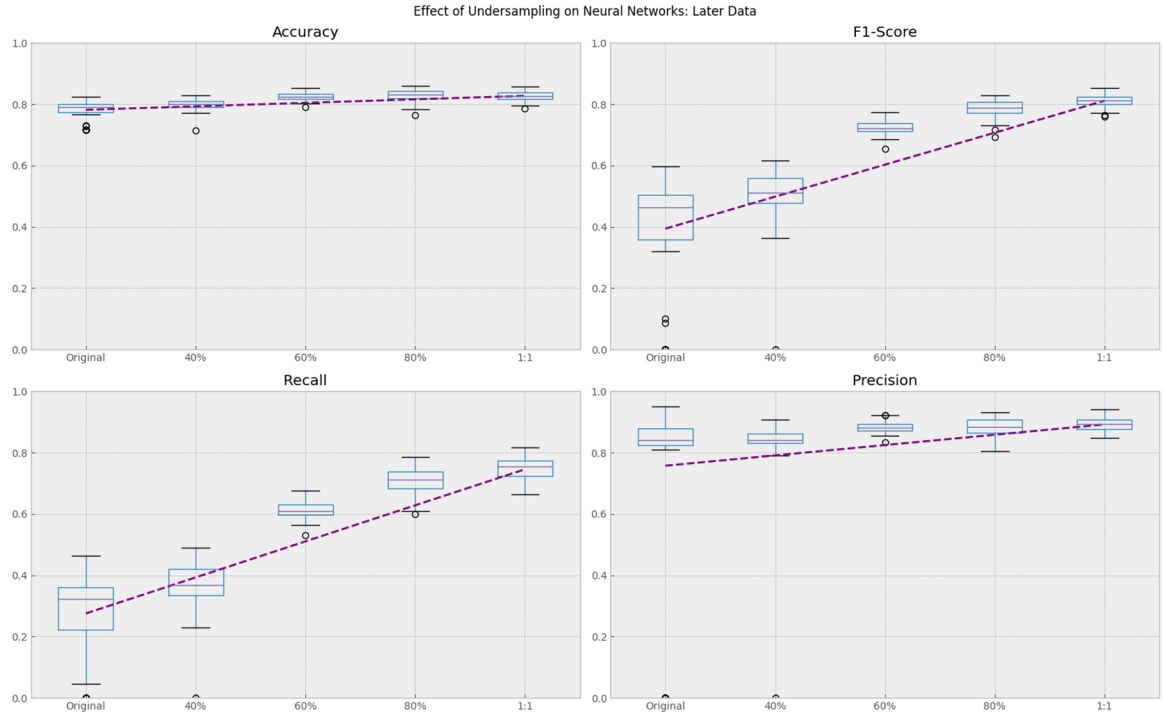


Figure 2: Random undersampling effect on NN performance considering varying ratios of imbalance.

Acknowledgments

This work was supported by the University of Louisville (UofL) and funded by NSF IUSE: EDU, Award #2335725. Nevertheless, any view, opinion, findings and conclusions or recommendations expressed in this material are those of the authors alone. Therefore, neither UofL or NSF does not accept any liability in regard thereto.

References

[1] V. Tinto, *Leaving college: Rethinking the causes and cures of student attrition*. University of Chicago press, 2012.

[2] V. Tinto and J. Cullen, "Dropout in higher education: A review and theoretical synthesis of recent research," *Office of Education (DHEW), Washington, D.C. Office of Planning, Budgeting, and Evaluation*, vol. 53, no. 9, p. 100, 1973.

[3] J. Bean and S. B. Eaton, "The psychology underlying successful retention practices," *Journal of College Student Retention: Research, Theory & Practice*, vol. 3, no. 1, pp. 73–89, 2001.

[4] C. P. Veenstra, E. L. Dey, and G. D. Herrin, "A model for freshman engineering retention." *Advances in Engineering Education*, vol. 1, no. 3, p. n3, 2009.

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[6] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[7] Y. Bengio, I. Goodfellow, and A. Courville, *Deep learning*. MIT press Cambridge, MA, USA, 2017, vol. 1.

[8] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[9] A. Fischer and C. Igel, "An introduction to restricted boltzmann machines," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings 17*. Springer, 2012, pp. 14–36.

[10] H. He and Y. Ma, *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[12] A. Tran, C. Zuniga-Navarrete, L. J. Segura, A. Dourado, X. Wang, and C. R. Bego, "Categorical variable coding for machine learning in engineering education," in *2024 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2024, pp. 1–5.