**RMAIR 2021 Presentation Proposal**

**Presenter**

Nathan Lindstedt, Data Scientist/Administrative Planning Specialist at Washington State University (Pullman, Washington)

**Category**

Research and Practice Presentation

**Title**

Predicting the Risk of Withdrawal in the First Year: Using a Machine Learning Approach to Forecasting Early Student Success

**Abstract**

Machine learning (ML) presents institutional researchers with new opportunities and challenges in addressing organizational concerns with student success. One such concern early on in students’ careers is retention past the first year. The use of predictive models trained on prior years’ data can help to inform administration, advisors, and other decision-makers on timely interventions that could reduce the number of dropouts and improve student success. This talk covers the origination of a ML model to predict first-year student retention at Washington State University. It provides an overview of the elements required and choices made in standing up this support tool.

**Summary**

This presentation is on the development of a predictive model for first-year students’ risk of withdrawal prior to the following academic year. It will define the population, detail the data and variables, describe the methodologies used, provide intuitions for how the ML algorithms work, outline how model predictions are classified, discuss the use and interpretability of global and local effects, cover the evaluation of model performance metrics, demonstrate how model predictions are communicated to decision-makers, and consider the limitations of these techniques.

The student success model primarily draws on student data and to a lesser extent on supplemental socioeconomic data to predict the probability that students will not persist after their first year. This is accomplished by first training the statistical model using historical data from prior year cohorts and then using that trained model to forecast the outcomes for the current year cohort.

The modeling strategy used is that of an ensemble model, which employs machine learning algorithms commonly used for binary outcomes including logistic regression, stochastic gradient descent, and multi-layer perceptron classification. The intuition behind this strategy is that each algorithm has its own inherent strengths and weaknesses that can be balanced to yield an aggregate model that is more robust to overfitting while also maintaining high predictive accuracy.

A voting classifier with weighted average probabilities is used alongside the ensemble model to combine the predictions of the included estimators to improve their generalizability. The weighted average of the predicted probabilities for the base models is calculated for each observation. The chosen weights determine the relative contribution of each algorithm to the average outcome.

Complexities associated with understanding why ML models produce a given set of results has caused them to be discussed as “black boxes.” In other words, while the practitioner knows the inputted data and the outputted predictions, it is more difficult to determine what factors explain the relationship between the inputs and outputs on a per-individual basis. Computational means have been established that enable us to peer inside these complex models irrespective of the algorithms used. The application of these methods permits for the identification of those unique features that contribute to students’ predicted probabilities of being retained after their first year.

The presentation will finish with an analysis of model performance metrics, a demonstration of the visual dashboard used for reporting results, and, finally, by talking about the possible blind spots of ML techniques.