**RMAIR 2021 Presentation Proposal**

**Presenter**

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

**Category**

Research and Practice Presentation

**Presentation Title**

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

**Presentation Abstract**

Machine learning (ML) presents institutional researchers with new opportunities and challenges in addressing concerns related to early student success. One such concern with early student success is the retention of first-year students. 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 presentation covers the development of a machine learning model predicting 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.

**Presentation Summary**

This presentation will provide information about the predictive model concerning first-year students’ risk of withdrawal prior to the following academic year. It will define the population of interest, detail the data and variables used, describe the data sources and methodologies used, provide intuitions for how the machine learning algorithms work, outline the process by which the model predictions are classified, discuss the use and interpretation of global and local effects, cover metrics of the evaluation of model performance, demonstrate how the model predictions can be communicated to decision-makers, and consider the limitations of these techniques.

The student success model primarily draws on student data and to a lesser degree supplemental socioeconomic data to predict the probability that a student will not persist into the next academic year. This is accomplished by first “training” the statistical model using historical data from prior year cohorts and then using that trained model to predict 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 against one another 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 in order to improve their generalizability. The weighted average of the predicted probabilities for the base models is calculated for each observation. The chosen weights thus determine the relative contribution of each algorithm to the average outcome.

The results of machine learning models have often been referred to as a “black box.” That is to say, while the practitioner knows the inputted data and the outputted predictions, it is more difficult to determine what factors explain the relationship between the inputted data and the outputted predictions on a per-individual basis. Numerical and computational techniques have now been developed that enable us to peer inside these complex models irrespective of the algorithms used. The application of Shapley additive explanations (or SHAP) permits for the localized interpretation of those features that contribute to students’ predicted probability of being retained after their first year.