**Beyond Early Warning Indicators**

* **Where does the data come from? What are the sources? Are these data available publicly?**

The data is collected from “[High School Longituidnal Study of 2009 (HSLS: 09)](https://nces.ed.gov/edat/help.aspx)”

Data Description: Micro study – around 26,000 students in 9th grade from 940 schools in 2009. Two level design:

* Step 1: Schools were selected at the national level
* Step 2: Around 30 students in each school were randomly selected among 9th graders.

The data results are publicly available (consent form needs to be signed in order to view the data.)

* **What are the machine learning methods used?**

1. LASSO
2. SVM
3. Logit

* **What are the predictors in the machine learning model? What is being predicted?**

The machine learning algorithm predicts high school dropouts. The independent variables used to predict initially included 1,700 variables. Using model selection algorithm, variables like GPA, age, math score, science score etc. were considered.

* **What are the econometric methods used?**

Economic model is introduced to consider budget constraints. An objective function is considered where, the number of students at risk of dropping out who are not treated plus those who are treated, each multiplied by the probability of dropping out given the treatment. The cost of the program in the budget constraint depends on the students who have been assigned to the treatment.

* **What is the intervention or treatment variable in the econometric method?**

The intervention variable is the treatment program given to the students who are categorized as weak.

* **How were the econometric methods & machine learning methods combined?**

The econometric method is used to obtain an efficient threshold value (keeping the budget constraint in consideration) which is used in the ML algorithm to maximize the recall rate.

* **What are the main strengths of the paper?**

1. The model is efficient theoretically and use of recall parameter along with accuracy is a good evaluation technique
2. The use of budget constraints makes the method achievable
3. Categorizing weak students using unsupervised ML algorithms is a good idea.

* **What are the two things the author could have done better?**

1. For identifying weak students, the author uses principal component analysis algorithms to decide what variables are the best predictors to use in a ML algorithm which classifies if a student is weak or not. Using these variables to generalize for the whole population may not be an efficient method.
2. The data can be collected quarterly to keep a track of progress and see if method is working. Instead of waiting for 4 years to see the result, the author could use the model to see if the students are dropping or not each year and evaluate the model.

* **Does the model make sense? Are the results believable?**

Though the model makes a lot of sense, the results are generalized to a large extent. Without proper evaluation technique, I would not believe them.

* **Is there anything you would do differently than the author?**

While considering the econometric methods, the author in order to form a closed-form expression, he assumes that either the students are weak or not, there are also cases when the students need help initially but once they are provided with help they get better and no longer would need any additional program. I would consider collecting data quarterly in order to evaluate the students and remove them from the program.