Early Warning System for At-Risk High School Students

NASCIO Award Category
Data Management, Analytics & Visualization
State of Illinois

Dr. Krishna Iyer, Chief Data Scientist
Krishna.Iyer@illinois.gov
Department of Innovation & Technology

Ali Sarchami, Data Engineer
Ali.Sarchami@illinois.gov
Department of Innovation & Technology

Debbie Stirling, Sr. Business Analyst
dstirling@isbe.net
Illinois State Board of Education

Initiation Date: February 2017
End Date: December 2017

EXECUTIVE SUMMARY

Illinois has long supported a state longitudinal data system (SLDS) that utilizes data to improve education. In 2015, the IT Division at the Illinois State Board of Education (ISBE) received approximately $7 million under an Institute of Education Sciences (IES) SLDS grant to accomplish the following:

- Link fiscal data to student outcome data within the ISBE data warehouse
- Establish school- and early childhood center-level accounting and reporting systems
- Standardize and automate local data collection processes
- Integrate local identity management and single sign-on with the Illinois longitudinal data system (ILDS) instructional supports
- Provide robust dashboard reporting and tools to Pre K-12 educators to identify learning gaps and improve student outcomes

The dashboard reporting system, known as Ed360, utilizes current and historical student enrollment, assessment, and attribute information to provide administrators and educators with data analytics on student performance and tools that support data-informed instructional practices. A component of Ed360 is to provide at-risk indicators on whether students are at-risk for not graduating high school and recommendations to guide interventions.

To develop at-risk indicators and a system to generate a predictive model, ISBE and the DoIT State Data Practice collaborated on a study that identifies major indicators based on student- and school-level data. A systematic stepwise correlation study and machine learning were employed to score students based on their middle and early high school academic performance to determine probabilistic drop out tendencies.

The outcome of the ISBE and DoIT collaboration is to provide Illinois administrators and educators with an early warning system that will identify at-risk students, recommend interventions, and monitor progress; as part of a statewide effort to prepare our students for college and career success.

CONCEPT

The biggest challenge facing educators in the State of Illinois is in putting students at risk of dropping out back on track to graduate high school. The sooner they can get an idea about the student's risk, the sooner they will be able to intervene. Graduating in a timely manner is essential for continued performance of the student, beyond high school, through college and any post-graduate work and into the workforce. Addressing the identified challenges sooner rather than later is important and in the best interest of the state as a market for job growth and the individual seeking a job.

The conventional approach for unraveling the challenge and making it workable is to use piecemeal statistical studies and run correlation studies to understand the impact of every contributor to the final outcome which is whether or not the student has graduated. On the other hand, in this initiative, advanced machine learning methods were utilized to train a model and enable classification of a student in terms of their risk of dropping out. This approach transcends the hypothesis testing methodology, as it considers all contributing factors at face value and then systematically sifts through the data to rank the variables based on their importance or contribution to the outcome. For this effort, tools available in the open source, based on Python and Scikit-Learn APIs were used to train and test the model. The workflow

for this methodology is highly iterative in nature, yet based on the scientific method. The process is illustrated in Figure 1. The key steps begin with (1) formulation of the business question; (2) identification of variables or data element; (3) data cleaning and preparation; (4) model building; (5) testing and verification and finally; (6) deployment of model into production.

The model prediction methodology for identifying dropouts at the end of high school is based on digesting a prior knowledge available on a student's history, from middle school through the first year of high school. The rationale for this approach is that educators should be able to use a guideline for determining what intervention programs may be necessary at the end of the first year of high school, so that the student can be placed back on track to graduate, detailed in Figure 2. For identification of data elements, a mind map is shown in Figure 3. This allows for a holistic approach to assembling all pertinent data elements from existing databases.

The approach is also scalable, and the study can be continuously improved year over year, by adding data elements that are newly added to systems. It is important to note at this point, that consideration is given equally to all the data elements identified, whether based on literature or other studies that have been shown to have an impact on a student's performance. The data sources are then verified for accuracy and integrity of the data before conducting some preliminary correlation studies. For model building, the dataset is randomly divided into training and testing sets; using the training set for building and optimizing and the testing set for verification. In this study, the student's middle school and first year of high school were used to model the outcome at the end of high school; the outcome being graduation. Care is taken to not overfit the model during the training phase, so that it becomes inflexible during the test phase or when deployed in production. Several algorithms are chosen and the one that fits the situation the best and lends itself to some interpretation is chosen during this phase. A significant benefit to this ever-evolving process is the ability to chain a series of models, beginning with elementary and middle school performance, that can eventually be chained to produce a master model for tracking and evaluating the risk at even earlier stages.

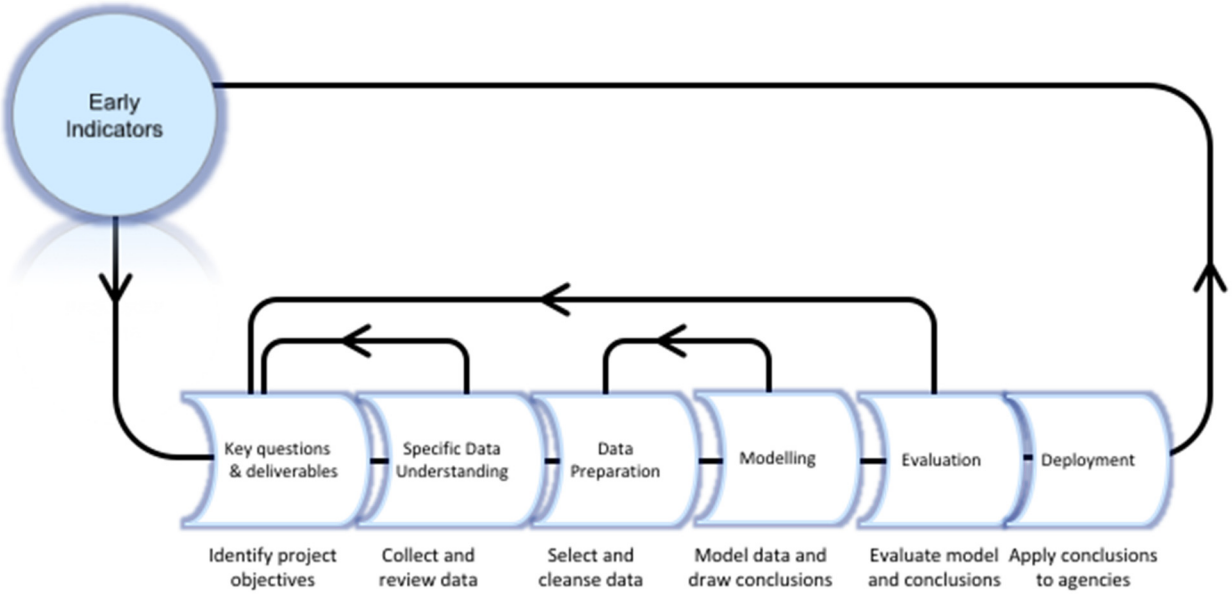

Figure 1: The Data Science model development cycle

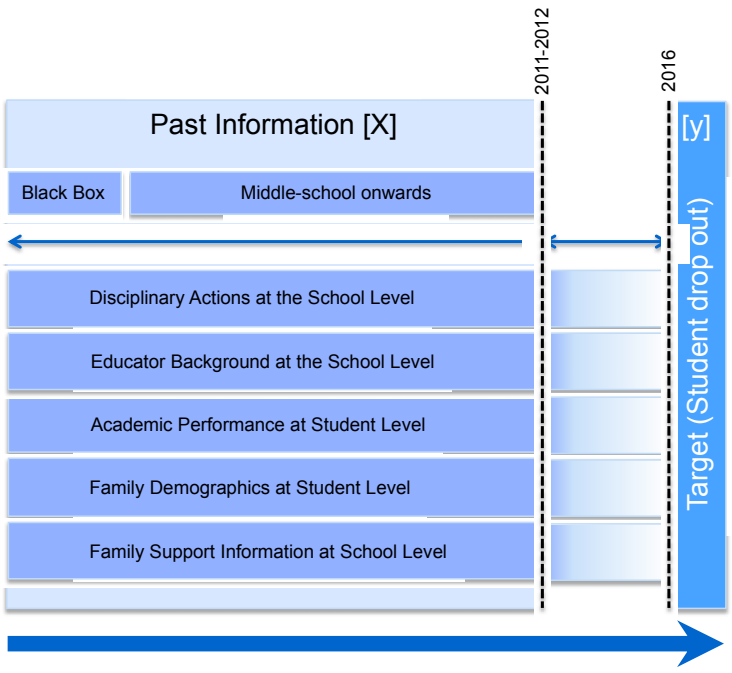

Figure 2: the prediction methodology

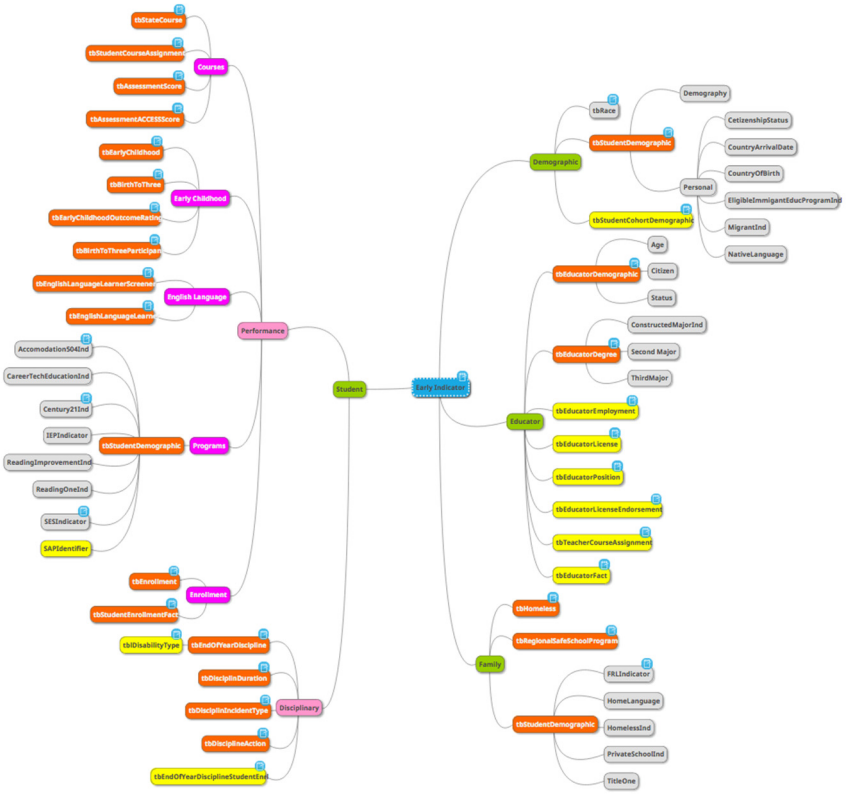

Figure 2: The mind-map for factors that influence student drop-outs.

SIGNIFICANCE

The most important aspects of this effort are the ability to digest a significant amount of information that is already available and then to be able to draw insights from the study that educators can act upon. The traditional approach relies on the educator's ability to determine which variables or factors might possibly impact a student's performance and to then complete a statistical study to validate or invalidate the hypothesis. Several statistical studies might have to be conducted in parallel and the end result may still be inconclusive. Machine learning circumvents this cumbersome challenge by taking all variables at the same time and allowing the data and the inherent patterns to dictate the importance of these variables.

It is worthwhile to mention a few of this effort's early, significant and actionable findings. Figure 3 shows a correlation plot of the binary label dropout (dropout = 0 and graduated = 1), against the variables used in this study. Behaviorally, students who participated well in school activities overall, including physical education and fine arts, tended to graduate. Whereas students who transferred between schools and were required to repeat a grade, tended to dropout. Figure 4 shows the most important variables of the model, a direct output of the algorithm used in this study overall. This successful model is being used today and will be repeated across the data variables and used to inform Board of Education Policy makers.

The algorithm also ranks a student by assigning a probability (between 0 and 1) that the student would drop-out. The higher the value, the higher the calculated risk for the student to drop out. The natural outcome of this leads to classifying a student using various categories: low, low-medium, high-medium and high. Educators benefit from this immensely as they can use this value to provide guidance, evaluate a student's needs and assign them to intervention methods and track their progress. This will also allow the state's initiatives to be measured in a timely manner for its effectiveness and impact.

The model is provided as an API and the data is refreshed daily in the respective systems to "score" a student. The model will also be tuned real-time if it falls below a desired accuracy. The results of this model will be made viewable through an existing statewide portal that educators are currently using.

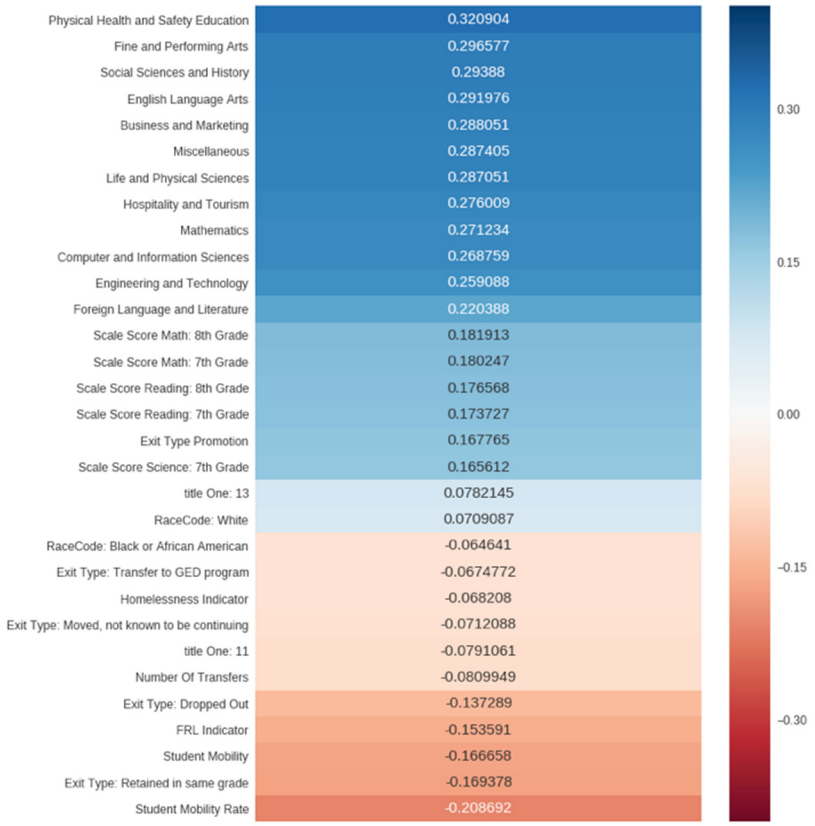

Figure 3: Correlation plot of student graduation.

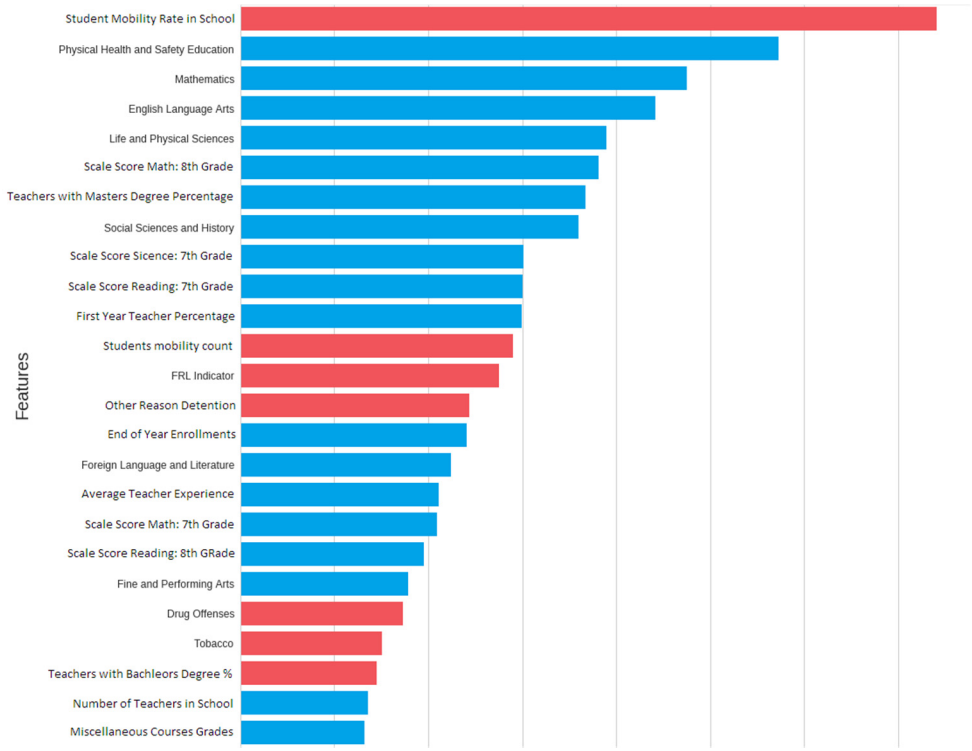

Figure 4: Most important variables plot color coded; favorable factors (blue), unfavorable factors (red).

	Actual	Probabilities	Risk
0	0	0.171116	Low Medium
1	0	0.565981	High Medium
2	0	0.679989	High Medium
3	0	0.239918	Low Medium
4	0	0.462101	Low Medium
5	0	0.091016	Low
6	1	0.855546	High
7	1	0.825125	High
8	1	0.602746	High Medium
9	1	0.877455	High

Figure 5: Model performance (actual vs probabilities) and student classification based on probability of dropping

IMPACT

With prior grant SLDS funding, ISBE used its student course history and grade data to establish a "Freshmen on Track" indicator for state public high school, based on whether students have earned at least five full-year course credits and no more than one semester "F" in a core course. This indicator, developed using nationally recognized research from the Consortium on Chicago School Research, provides a key predictor of high school success enabling active interventions at the school level.

Recent research indicated that adding short-term metric changes of key at-risk indicators (during and across multiple school years) to a predictive model can increase the utility of using an early warning system to measure student improvement and evaluate interventions. Research also established that predictive indicators may work for one district but not in another district. For this reason, exploring local interventions to generate localized predictive change indicators will increase the utility of the early warning system to help the local school staff make data-informed intervention decisions and establish a set of local at-risk practices. For example, how best to provide individual or small group interventions and how to monitor and measure progress.

Educators are less interested in the risk level and more interested in knowing which interventions will be more useful for a specific student. In addition, training the model to recommend interventions based on a student's likelihood of responding to that intervention will significantly influence educators' interest in using the system. More importantly, providing a tool that can accurately match an intervention to a student's need will significantly impact dropout prevention.

Lastly, based on a 2013 study conducted by Alliance for Excellent Education, there are numerous economic benefits that can be realized from an early warnings system for the State of Illinois. A projected increase of 33,000 graduates in high school can enable 3,150 new jobs, create $376 million in additional earnings, $279 million increase in annual spending and contribute $518 million to gross state product. The Illinois State Board of Education is on track to expand the production model to include additional data and predictors.