Predicting Student Risks Through Longitudinal Analysis

Ashay Tamhane
IBM Research - India
ashayuda@buffalo.edu

Shajith Ikbal
IBM Research - India
shajmoha@in.ibm.com

Bikram Sengupta
IBM Research - India
bsengupt@in.ibm.com

Mayuri Duggirala*
Tata Research Development & Design Centre, Pune, India
mayuri.duggirala@tcs.com

James Appleton
Gwinnett County Public Schools, GA, USA
jim_appleton@gwinnett.k12.ga.us

ABSTRACT

Poor academic performance in K-12 is often a precursor to unsatisfactory educational outcomes such as dropout, which are associated with significant personal and social costs. Hence, it is important to be able to predict students at risk of poor performance, so that the right personalized intervention plans can be initiated. In this paper, we report on a large-scale study to identify students at risk of not meeting acceptable levels of performance in one state-level and one national standardized assessment in Grade 8 of a major US school district. An important highlight of our study is its scale - both in terms of the number of students included, the number of years and the number of features, which provide a very solid grounding to the research. We report on our experience with handling the scale and complexity of data, and on the relative performance of various machine learning techniques we used for building predictive models. Our results demonstrate that it is possible to predict students at-risk of poor assessment performance with a high degree of accuracy, and to do so well in advance. These insights can be used to pro-actively initiate personalized intervention programs and improve the chances of student success.

Keywords

education; educational data mining; risk prediction; longitudinal data analysis

1. INTRODUCTION

One of the primary goals of any education system is to equip students with the knowledge and skills needed to transition to successful career pathways. How effectively education systems around the world are able to meet this goal acts as a major determinant of economic and social progress. In particular, K-12 reflects the most critical phase of an individual's lifelong learning, during which the opportunities for a successful future need to be created and nurtured. It is in recognition of this fact that increasingly, educational reforms focused around frequent and standardized testing of K-12 students are gaining acceptance as a way to monitor performance and progress at an individual and institutional level [3, 4]. For many of these assessments, a passing score is required to meet state or district standards for progression to the next grade level. Poor academic performance in such standardized assessments can thus lead to unfavorable educational outcomes such as grade retention, and when not adequately addressed, it can eventually trigger dropout or sub-optimal career pathways, which are associated with significant personal and social costs. For example, in the United States, nearly 7,000 high school students drop out of school each day; if the students who dropped out of the class of 2011 had graduated, the nation's economy would have benefited from $154 billion in additional income over the course of their lifetimes [11]. Hence, it is important to identify students at risk of poor performance in major standardized tests, and to do so well in advance, so that the right personalized intervention plans can be initiated to improve performance.

Traditionally, K-12 educators (e.g. class teachers) have relied on recent academic results of a student (e.g. in formative tests in the current grade) along with an educator's general intuition gleaned from teaching similar students in the past, to determine if the student might be at risk of poor performance in an upcoming assessment. This makes the process overly reliant on an educator's experience level, there is no objective quantification of the level of risk, and the dependence on recent data to make a prediction for the academic year may often not leave enough time to apply the right level of intervention to adequately improve performance. However, with the digitization of school records and rapid uptake of digital tools for teaching and learning, various aspects of a student's longitudinal journey through K-12 are now captured and persisted in digital form. This offers a rich repository of data that can be analyzed to detect patterns associated with unsatisfactory educational outcomes, derived from thousands of students who have progressed through the system over the years, and taking into account a holistic view of a student in terms of both academic history over a period of time, as well as other non-academic attributes (e.g. related to attendance, demographics, behavior etc.) that may influence academic performance.

*This work was done while the author was at IBM Research-India.

KDD'14, August 24–27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2623330.2623355.

The work reported in this paper has been motivated by the need to develop scientific, robust predictive models for at-risk students in a large K-12 school district in the US. An increasing number of school districts now have substantial volumes of historical data that have resulted from digitization efforts over the last decade, spurred by legislation that mandates implementation of longitudinal data systems and usage of such data to improve instruction [4]. The district in question - Gwinnett County Public Schools (also referred to as GCPS in this paper), based near Atlanta, Georgia - is the largest school system in the state of Georgia, serving more than 168,000 students across 132 schools in 2013-14 [1]. The availability of a district-wide common data system allows the progress of students to be tracked longitudinally across its 77 elementary, 26 middle and 19 high schools, besides 4 charter schools and 6 other special schools. In the course of this K-12 journey, students take a variety of standardized assessments, both state-level and national-level, and detailed performance data in these tests are persisted within the GCPS data warehouse. In addition, a variety of other data about the student such as enrollment history, various demographic indicators, discipline/behavior etc. are also stored. Over the years, such data on more than 200,000 students have been made available in the GCPS warehouse, which presents both a challenge in terms of its sheer volume, variety and complexity, and also a tremendous opportunity to develop sophisticated models of performance that can be used to improve teaching and learning programs.

Within this overall context, our specific objective was to develop predictive models to identify at-risk students in Grade 8, for two disciplines mathematics and science, in two standardized tests – a state-level assessment called Criterion Referenced Competency Test (CRCT) and a national assessment called Iowa Test of Basic Skills (ITBS). Grade 8 was specifically chosen because of its significance in the K-12 journey – as the final year in middle school, preceding the first year of high school. The specific assessments and disciplines were selected based on their importance to GCPS. The research and experimentation were driven by the following questions: (i) how accurately (and using which feature sets) can we predict students who fail Grade 8 CRCT and ITBS Mathematics and Science tests? and (ii) how early (by which grade), can we make these predictions with reasonable accuracy? The first question is important because unless at-risk students and not-at-risk students are differentiated with reasonable accuracy, significant resources may be expended on misaligned interventions for these student cohorts, without ultimately achieving the desired improvement in success rates. The importance of the second question stems from the fact that the earlier we are able to accurately identify an at-risk student, the more the time available to the school to apply interventions to reduce the risk of failure.

Our models are built based on the longitudinal data derived from the GCPS data warehouse on students who have appeared for CRCT and ITBS Grade 8 assessments in mathematics and science over the last several years, which numbered 58,361 and 43,306 students respectively for these two tests. Mean imputation was used to address missing values in the longitudinal trajectories of the students, and the models were developed employing the techniques of Logistics Regression, Naive Bayes and Decision Tree, using the IBM SPSS Modeler [2] and Weka [14]. The data attributes used for deriving the features of the predictive models relate to performance in various assessments in mathematics, science and related disciplines, several demographic indicators (e.g. gender, ethnicity, free or reduced meal eligibility, special education/gifted) and behavioral indicators (e.g. suspensions).

Our key findings may be summarized as follows:

- Our results indicate the it is possible to predict students at-risk of failing Grade 8 CRCT and ITBS assessments in Mathematics and Science with a high level of accuracy and balance between differentiating at-risk students and not-at-risk students. Of the algorithms, logistic regression gave the best results overall.
- We also observe that it is possible to predict success/failure in the Grade 8 assessments as early as Grades 4 and 5, with data from subcequent grades further improving the effectiveness of the predictions
- In terms of features, longitudinal test scores from earlier grades emerge as a strong predictor, while the set of demographic indicators taken together, are able to achieve reasonably high prediction performance as well.

The rest of the paper is structured as follows: in Section 2, we discuss related work. A detailed description of the data sets used for the analysis and its processing and feature engineering are presented in Section 3. Section 4 discusses the prediction and experimental set-up, including the data, evaluation metric and the features used. Section 5 presents the experimental results, while Section 6 has a discussion on some of the insights from the results. Finally, Section 7 outlines future work directions, and Section 8 concludes the paper.

2. RELATED WORK

There is a wealth of research available on the factors that contribute to student risk of academic performance and dropout. With the increasing availability of education related longitudinal data within K-12 and Higher Education institutions, this line of research is now becoming increasingly data-driven and evidence-based. Our study adds to this body of work, from which we review a few selected papers below to put our research in context.

Previous research has studied student risk in terms of factors such as low socioeconomic status, living in a single parent home, changing schools at non-traditional times, below average grades at specific school levels, being held back in school through grade retention, having older siblings who left high school before completion and negative peer pressure [19]. Other external factors influencing student risk, such as teacher engagement [7] and smaller class size [20] have also been studied. Several typologies have been developed for describing students at risk. Fortin et al. [12] classified at-risk students into four different subgroups: the covert behavior type, the uninterested in school type, the school and social adjustment difficulties type and the depressive type, thereby highlighting the importance of school, family and personal factors in the emergence of risk. Summarizing twenty–five years of research on influences of dropout, Rumberger and Lim [18] delineated rigorous empirical results into a typology of predictor categories. These categories included: 1) individual–level factors which consisted of background, attitudes, behaviors, and performance, as well as 2) influential, but perhaps less commonly considered family–, school–

and community–level variables. At the student–level, background was focused upon demographic and health variables as well as past school experiences and performance. These influences impacted attitudes which were believed to impact behaviors and, in turn, academic performance. Such findings have influenced our selection of predictors of student performance in our study, where we have leveraged a variety of demographic and behavioral data of students, in addition to their historical academic performance record, to predict future performance.

In recent years, due to the digitization of school records and availability of instrumented digital learning environments, a wealth of student related information is now available in data warehouses, providing a fertile ground for educational data mining research. The Signals project at Purdue University [5] is a notable effort in this direction. Signals combines online academic behavior of students, such as whether they opened or completed assignments and exercises, with a student's academic performance such as standardized test scores, high school GPA and current grades, to infer the risk level of the student in the course. Other online systems make use of students' social engagement with their peers in the context of work, in addition to content-access patterns, to identify disengaged learners who can potentially be at-risk [10]. Some key variables used in predicting students at risk in online courses typically include total number of discussion messages posted, total number of mail messages sent, and total number of assessments completed. However, these models are limited in their generalizability as they focus on fully online courses alone [17]. In our current study, data from digital learning environments was not available for analysis, and we intend to extend our work with such data when available, to investigate how it influences the predictive power of our models.

In a recent study, Chen and Elliott [8] have reviewed previous research which highlights the role of early identification of students at risk. Specifically, their study found that any student who had passed the key predictor course module seemed more likely to progress to the second year successfully. Similarly, any student who failed the predictor module was more likely to repeat some of the first year modules, repeat the entire first year, or, even fail the course, i.e., drop out from a course. This pattern was valid for all of the records of the five consecutive academic years considered for the study. Our study adopts a similar approach where we attempt to examine performance (test scores) in the current grade based on test scores in previous grades, both at the micro (test strand) level and the macro (subtest) level. We also demonstrate how risks for Grade 8 standardized tests can be predicted with a reasonable degree of accuracy a few years in advance.

At the high school and undergraduate level, Hershkovitz et al [15] used a learning graph which represents the student's learning over time and was developed using a knowledge–estimation model. This model infers the degree of learning that occurs at specific moments rather than the student's knowledge at those moments. It showed substantially better student–level cross validated prediction of student's future learning than previous approaches. However, the study used a limited student sample for its research (n=181 undergraduate and high school students) thus limiting the generalizability of its conclusions. In a study on students in tertiary education by Gray, McGuinness and Owende [13], models of academic performance were found to achieve good predictive accuracy when younger students and mature students were modeled separately. Further, students with missing data were removed from analysis and the study was carried out with students having complete data. Missing data is an unavoidable characteristic of any large real-life data set, and hence we addressed the issue in our research by imputing the missing data with the mean, rather than only selecting a small number of students with complete data on a few features. Using Bayesian networks, Vihavainen et al. [21] demonstrated that students with a higher likelihood of failing their mathematics course could be detected at an early phase of their studies using data on their programming behavior. The sample comprised a large number of programming snapshots but only from 58 first year computer science students. In a recent study, Erdos et al [9] studied the extent to which first language (L1) predictors can be used to predict risk for French (L2) reading and language learning. Analyses of 86 kindergarten children revealed factors such as phonological awareness, phonological access and letter–sound knowledge in L1 were significant predictors of risk for reading difficulties in L2. Similar research has examined the relationship between oral reading fluency and success on state standardized assessments [16]. Our work builds on such examples of the use of data mining and machine learning techniques to predict academic risks, but is characterized by very large real-life data sets e.g. models built from more than 58,000 students drawn from 132 schools, and leveraging 342 features, in the case of ITBS success prediction. This gives a very broad data and evidence foundation to the study, which contributes to the robustness and generalizabilty of the results.

A review of research considered the predictive properties of 110 indicators of dropping out of high school across 36 studies and classified these on several dimensions [6]. Of these dimensions, the authors argued for sensitivity (the proportion of accurately identified future dropouts (or non–graduates) – "true positives") and specificity (the proportion of accurately identified future graduates – "true negatives") as the most important qualities of an indicator. We have adopted these metrics when evaluating the effectiveness of our predictive models.

3. DATA DESCRIPTION

Gwinnett County Public Schools (GCPS) is one of the largest school systems in the US, consisting of 132 schools and serving more than 168,000 students at present. A variety of data related to students, teachers, as well as learning and assessment activities is collected from each of the constituent schools and collated into a central data warehouse, offering a rich repository that can be mined for insights. A snapshot of this data warehouse was made available to us, which we use as the source of data for our analysis. The data snapshot consists of over a hundred tables storing various types of data including student, teacher, school, and test (assessment) data. These four broad categories are shown in Fig.1. The most important category of the available data (for the purpose of this paper) is the 'Students' category, which is a group of tables storing student data like enrollment, demography, test performance, course history and their performance in various national and state level tests they have taken so far. For the privacy of students and teachers whose data is captured, their personal infor-

mation was anonymized by replacing it with hypothetical names and addresses before the data set was made available for this study.

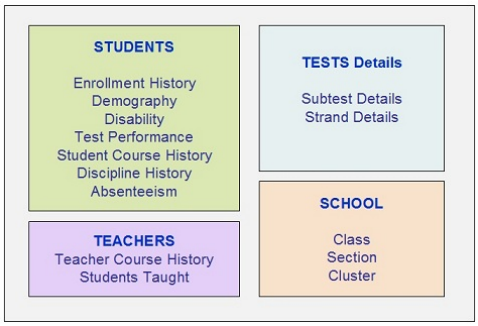

Figure 1: An Overview of the Available Data

3.1 Background On Tests and Defining Risk

It is important to understand the structure of the tests, as it will help us explain the features we use, our definition of risk, as well as understand some of the missing data challenges later in the section. Every "test" or assessment has several "sub-tests". For example, the Criterion References Competency Test (CRCT) is a test which has several sub-tests like mathematics, science and literature. Every sub-test further consists of different "strands". For example, mathematics has strands like algebra and geometry. All the test scores in the data are available both at sub-test and strand levels. Note that some tests only have sub-tests and do not have strands (for example, CogAT).

Since 2005-06, CRCT has been Georgia's annual assessment for determining how well students have acquired the skills and knowledge described in the Georgia Performance Standards (GPS). The CRCTs are currently taken across grades 3-8. In 2012-13, the CRCTs transitioned to include Common Core (CC) based items and into the (CCGPS/GPS) version. For our study, we consider the GPS version of CRCT so as to have a comparable sample of students.

Iowa Test of Basic Skills (ITBS) provides the opportunity for comparison with a nationally representative group of peers. Through ITBS, GCPS assesses students in grades 3, 5, and 8 in the areas of reading, written expression, vocabulary, conventions of writing, mathematics, science, and social studies.

The Cognitive Abilities Test (CogAT) is an assessment designed to measure acquired reasoning abilities in students, covering areas most linked to academic success in K-12 (verbal, quantitative, non-verbal). For this reason, we used CogAT results as predictors for performance in other assessments like CRCT and ITBS.

CRCTs measure performance based on predefined ranges of scores. CRCTs are structured so that the they range from 650–900s. Scores at or above 850 indicate exceeding standards, those at or above 800 indicate performance that meets test standards, where as those below 800 indicate below test standards performance. We consider a student to be at risk, if he scores below 800 in CRCT in the subtest of interest (mathematics or science). ITBS proves percentile ranks (PR) – which indicate rank within a distribution and when used with a distribution of nationally–representative peers, support inferences of relative national performance. For our analysis, we use the national percentile ranks as an indicator of performance in the test. Being a norm-referenced test, there are no predefined performance ranges unlike CRCT. Our subject matter expert from GCPS recommended a threshold of 25 percentile to mark the risk for ITBS. We derive our target risk variables by putting corresponding thresholds on grade 8 CRCT and ITBS scores as we intend to predict risk of students performing poorly in grade 8.

3.2 Missing Data

All together, the data warehouse contains historical data of over 200,000 students. However, the data is not consistent both within and across data categories. For example, a student's performance in CRCT may be available for grades 4-8. However, ITBS scores might be available only for grades 3 and 5. There may be several reasons for such missing data. Students change counties and schools multiple times during the K–12 period. For the period when students are not part of the county, the data of their performance may not be available. A student may have not taken certain assessments. Education standards change from time to time, and so do the tests associated with them, as discussed for the case of CRCT test. Finally, data warehouse snapshot available to us may not have the complete historical data available to the school.

We now present the data processing and feature engineering pipeline which we follow to create data sets suitable for our prediction task, considering the missing data challenges.

3.3 Data Processing and Feature Engineering

We divide the task of extracting useful data from the data warehouse and converting it into usable datasets into three main phases.

3.3.1 Creation of Merged Data

As described earlier, the data warehouse contains over hundred tables. Unfortunately, the Student data which we are interested in is scattered across several tables. The first phase consists of creating a data set with features extracted and engineered from various tables and columns in the data warehouse. This phase (Fig.2) involves three main stages.

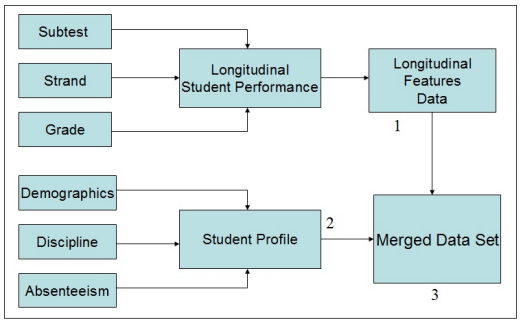

Figure 2: First Phase of Data Processing

Firstly, we create Longitudinal Features Data. This stage involves merging the different tables storing data on sub–tests and strands, as well as mapping performances to the corresponding grades in which students took the tests. This results in intermediate Longitudinal Performance Data, which

contains about 19 million records as it consists of all the performances of every student. We use SPSS Modeler streams [2] to restructure the data so as to generate features for each student from the Longitudinal Performance Data. Currently, we only consider CRCT, ITBS and CogAt test features in this stage. Note that these features include both the sub-test level and the strand level scores, when applicable. The Longitudinal Features Data contains data for 158,282 students and altogether 516 different columns each containing a specific test score. However, it is important to note that large amount of entries in this data are missing and we shall discuss how we handle this in the final phase of data processing.

The second stage involves creation of the Student Profile which contain features engineered from the demographics, discipline history and absenteeism history. While the demographic features like Gender, Ethnicity, Free Meal, Gifted and Special Education Needs are available straight from the warehouse tables, features like 'Number of Discipline Incidents', 'Number of Suspensions' and 'Number of Absent Days' were created after aggregating discipline incidents and absenteeisms reported.

In the third stage, we simply merge the Longitudinal Features and the Student Profile Features, resulting in a Merged Data set which contains several missing rows and columns from the Longitudinal Features Data.

3.3.2 Creation of Target Variable Dependent Data

The second phase involves creating target variable specific data sets from the Merged Data set obtained from the first phase of data processing. Firstly, we select only those students who have no missing values for the target variable. For example, when the target variable is CRCT 8th Grade Math sub–test, we discard all those students whose data for this target variable is missing. After this step, we discard features which have more than 80% missing values. This results in 58,707 students with 342 features for ITBS test, and 43,310 students with 282 features for the CRCT test. Note that though some of the columns may contain as low as 20% values, the fact that our data covers a very large number of students ensures that at least 11,741 samples exist for each feature in the ITBS set and 8,662 samples for the CRCT set. Nonetheless, since most of the standard classifiers do not work with missing data set, we impute the missing values in the third phase.

3.3.3 Imputation of Missing Features

To make our system robust, we handle missing data which routinely exists in data warehouses of the scale similar to that of GCPS. Moreover, none of the students contain all the features in our data set. Therefore, creation of a complete data set containing features ranging over all the different grades is not possible. For the sake of simplicity in terms of method, we impute the missing values using mean value imputation. This is a common imputation technique which simply replaces each missing value with the mean of the feature. However, many advanced imputation techniques exist and exploring them for our system remains a work to be done in future.

4. RISK PREDICTION AND EXPERIMENTAL SETUP

4.1 Data

We created data sets for 3 different tasks as below to demonstrate that our approach is generalizable to various student risk prediction tasks:

1. Predicting risk of poor performance in CRCT grade 8 Mathematics subtest
2. Predicting risk of poor performance in CRCT grade 8 Science subtest, and
3. Predicting risk of poor performance in ITBS grade 8 Mathematics subtest.

During the actual experimentation, we excluded all the features related to grade 8 or above, to account for the fact that prediction will be done only using grade 7th or below features. As discussed in earlier sections, we define our risk (derive our target risk labels) using thresholds on CRCT and ITBS grade 8 Math/Science scores. Thresholds used are 800 for CRCT and the 25 percentile for ITBS. We address the problem of predicting risk by transforming this into a binary classification problem, where at-risk students are labeled as positive samples and non-risk students labeled as negative samples. As a preprocessing step, we standardized all the features to mean value of 0 and standard deviation of 1.

In our experiments, we used 5-fold cross validation settings to evaluate the accuracy of predictions made in various tasks. In this setting, the entire data set for a given task is divided into 5 equal parts. Prediction experiment is repeated 5 times, each time keeping one of the 5 parts as evaluation data and the rest as training data to build prediction model. In the end evaluation results on all 5 parts are accumulated together to compute the overall prediction accuracy.

4.2 Prediction

We used binary classifiers as risk prediction models. Specifically we used various implementations of classifiers in SPSS [2] and WEKA [14] with their default settings, such as logistic regression, naive bayes, decision tree and decision table. All the logitudinal features available per student are used as set of input features to predict a binary output namely, whether a student is at risk in various tasks listed in previous section.

4.3 Evaluation Metric

The ITBS data set contains 58,361 samples containing 15.3% positive samples and 84.7% negative samples. This shows the between-class skew that exists as students with such extreme risk are naturally rare to find. Similar is the case for the CRCT data set – which contains 43,036 students including 10.7% positive and 89.3% negative samples. This heavy skew in data has important implications on the evaluation criteria.

As described earlier, student risk prediction is effectively a binary classification task aimed at categorizing students into two groups: 1) *risk* and 2) *no-risk*, depending upon whether a student is likely to perform poorly in an important examination or not. Since the number of students falling into

risk category is typically smaller, a simple classifier where all the students are assigned to *no-risk* class is expected to give high classification accuracy, in spite of the fact that it would result in very poor (0%) *risk* class prediction accuracy. For this reason, in this paper, we have used *receiver operating characteristic (ROC) curve* to measure the prediction performance. ROC curves show a trade-off achieved between true positive rate of the *risk* category (sensitivity) and false positive rate of the *no-risk* category (specificity) for various classification thresholds applied on classifier class probability outputs (see Figure 3 for a sample ROC curve plot). True positive (TP) rate is the percentage of actual *risk* students correctly categorized into *risk* category. False positive (FP) rate is the percentage of actual *no-risk* students wrongly categorized into *risk* category. TP rate is expected to be high for low class probability thresholds. However, this is expected to also result in high FP rate. On the other hand a high probability threshold is expected to result in lower TP and FP rates. An example ROC curve obtained for our task can be found in Figure 3.

Typically in practical systems, depending upon the task requirement, a best class probability threshold is chosen to achieve one of TP rate or FP rate requirements. However, in this paper, we use area under ROC curve (AUC-ROC) as a measure of overall prediction performance. Increase in AUC-ROC could be achieved by pushing the ROC curve shown in Figure upward left, which in turn means an improved TP and FP rates for a given class probability threshold. AUC-ROC typically varies from 0.5 to 1.0. A simple classifier choosing one of the class labels (say the majority class label) as class output would achieve an AUC-ROC value of 0.5.

4.4 Features Used

At the end of the data processing (as described in Section 3.3), total number of features available for CRCT risk prediction is 235 and for ITBS risk prediction is 280. Based on the description given in Section 3 , these features can be grouped into categories based on their type. Specifically we have grouped these features into following broad categories:

- Scores: scores obtained by students in their past grades from 1 to 7 in various tests including CRCT, ITBS and CogAt.
- Demography: student demography information such as gender, ethnicity, free meal, gifted and special education.
- Behavioral: information recorded about student behavior such as number of absent days, number of suspensions and number of discipline incidents.

We further grouped the score features into categories based on the grade at which the score is obtained (such as grade 1, grade 2, ..., grade 7) and the subject in which the score is obtained (such as maths, science, literature and others).

5. EXPERIMENTS AND RESULTS

In this section we discuss results of experiments performed to evaluate the student risk prediction system developed for GCPS. These experiments are aimed at two goals: 1) to measure risk prediction performance (described in Section 5.1), and 2) to analyze the importance of various aspects of student data and prediction in order to derive insights that would be of potential value to the educational institutions. Specifically, in such analysis we aim to explore two different aspects namely: data and prediction that might be of potential interest to educational institutions such as: 1) Feature importance (described in Section 5.2, and 2) Early prediction of the risk (described in Section 5.3).

5.1 Risk Prediction Performance

Table 2 shows a comparison of AUC-ROC values obtained using various classifiers for risk predictions in CRCT 8th grade Mathematics, CRCT 8th grade Science and ITBS 8th grade Mathematics. We used various classifier implementations in SPSS [2] and Weka [14] with their default settings. For all the tasks, a relatively simpler logistic regression model is able to consistently achieve best prediction accuracy. Hence, we have chosen to use logistic regression model in all our further experiments. As discussed in Section 4.3, ROC curve is a trade of between true positive (TP) and false positive (FP) rates. An example ROC curve obtained using logistic regression model for CRCT 8th grade Mathematics risk prediction task is shown in Figure 3. Typically, educational institutions want to keep the incorrect prediction of true risk students as low as possible. This in turn means a specification of required TP rate for the prediction task. A preferable TP rate for GCPS we are working with is 90%. Table 1 shows FP rate we could achieve using the logistic regression classifier for a minimum TP rate of 90% for various prediction tasks.

Task	Probability Threshold	True Positive TP, in %	False Positive FP, in %
CRCT 8th Mathematics	0.06	90.5	23.8
CRCT 8th Science	0.18	90.0	24.7
ITBS 8th Mathematics	0.1	90.7	28.8

Table 1: Probability threshold used to achieve minimum true positive (TP) of 90% in risk prediction. Actual values of TP and FP achieved for the thresholds used are also given.

Classifier	CRCT 8th Grade Mathematics	CRCT 8th Grade Science	ITBS 8th Grade Mathematics
Naive Bayes	0.744	0.739	0.702
Decision Tree	0.822	0.774	0.766
Decision Table	0.933	0.902	0.893
Logistic Regression	0.924	0.907	0.896

Table 2: Comparison of classifier performances for risk prediction in CRCT 8th grade Mathematics, CRCT 8th grade Science and ITBS 8th grade Mathematics.

5.2 Feature Importance

In Section 5.1 we used all available student features in order to predict risk with highest accuracy. However, educational institutions are typically interested not just in the overall prediction but also in exploring relative contribution of various features. Particularly, they are interested in finding out a subset of features that are more indicative of potential risk than others. For this purpose, we have performed a feature-wise analysis of ability to predict.

Table 3 shows a performance comparison of the various feature groups described in Section 4.4. Among the broader feature groups, score features achieve the best prediction accuracy with performance almost to similar that of the entire

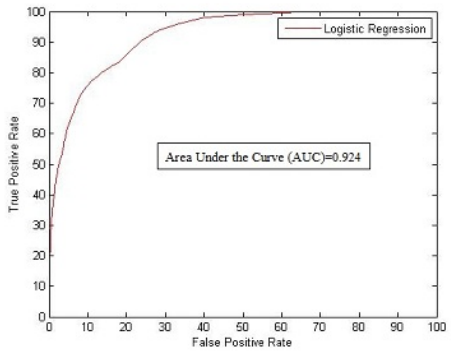

Figure 3: ROC curve for risk prediction using logistic regression for CRCT 8th grade Mathematics

feature set. This strengthens the fact that the past scores are largely indicative of the future test performance. Interestingly, the small set of demography features (5 features) are also able to achieve a reasonably good accuracy, pointing to the fact that specific patterns of demography features are typically indicative of a potential risk. However, behavioral features do not seem to correlate well with the future risk. In fact, further examination of the behavioral features show that only a few number of students are recorded for behavioral issues. Among individual demography features, ethnicity and special education needs achieve best accuracies. Table 3 also shows a comparison of importance of past subject-wise score features in predicting risk for 8th grade ITBS Mathematics. Clearly the past Mathematics scores are able to make better prediction than the other subject scores.

Feature Type	CRCT 8th Grade Mathematics	ITBS 8th Grade Mathematics
All Features	0.924	0.896
All Scores	0.902	0.882
All Demographics	0.866	0.814
All Behavioral	0.576	0.559
Scores - Maths	-	0.871
Scores - Science	-	0.828
Scores - Language	-	0.846
Scores - Others	-	0.829
Demography - Gender	0.547	0.537
Demography - Ethnicity	0.660	0.668
Demography - Gifted	0.622	0.630
Demography - Free Meal	0.646	0.640
Demography - Special Education Needs	0.721	0.637
Behavioral - Absence	0.537	0.542
Behavioral - Suspensions	0.588	0.578
Behavioral - Incidents Reported	0.583	0.569

Table 3: Comparison of feature level risk prediction performance for CRCT 8th grade Mathematics and ITBS 8th grade Mathematics.

5.3 Early Prediction of the Risk

Educational institutions are typically interested in predicting students-at-risk as early as possible so that they can pro-actively adopt early intervention plans to prevent potential failures. For this purpose, we investigated prediction of potential risk at 8th grade using only the lower grade features. Figures 4 and 5 show results of such prediction. In fact, in these figures, risk prediction performance using aggregated features as well as individual features are shown together for comparison. In case of aggregated features, all the score features from grades equal to or less than that mentioned on the x-axis are used for prediction whereas in case of individual features, scores only from grades mentioned in x-axis are used. The results show that we are able to make reasonably accurate predictions (ROC AUC of more than 0.8) as early as 4th grade for 8th grade performance. Also the prediction accuracy improves with the incremental aggregation of score features from lower grades. Individual scores from the recent past are also good predictors of the risk. However, the aggregated features are always able to predict with higher accurately than the individual features.

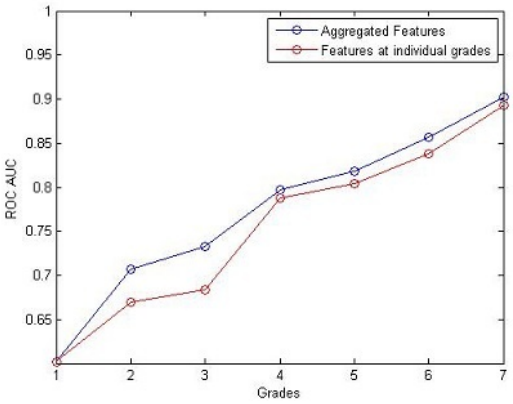

Figure 4: Early risk prediction performance for CRCT 8th grade Mathematics. In case of aggregated features all the score features equal to or less the grade mentioned in x-axis are used. For comparison, performance using individual grade scores is also given which uses just the scores from grades mentioned in the x-axis.

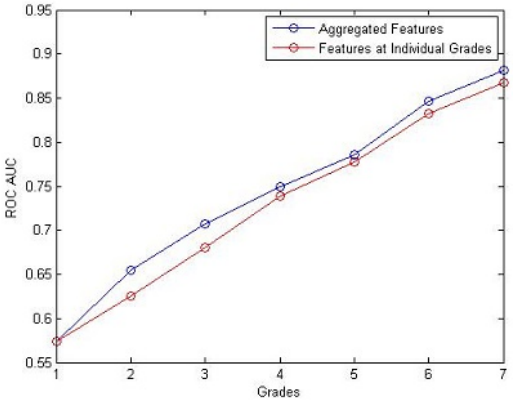

Figure 5: Early risk prediction performance for ITBS 8th grade Mathematics. In the case of aggregated features, all the score features equal to or less than the grade mentioned on the x-axis are used. For comparison, performance using individual grade scores is also given which uses just the scores from grades mentioned on the x-axis.

6. OBSERVATIONS AND INSIGHTS

In this section we discuss some of the insights derived from our experimental results that have potentially valuable implications for educators and industry practitioners.

First, our results showed that a student's risk of poor performance can be predicted with reasonable accuracy (ROC AUC of more than 0.8) in as early as grade 5. Such early prediction can allow teachers sufficient time and fewer resources to take remedial actions in a student's learning path. Additionally, we see that as the grade of prediction increases, the ROC AUC steadily increases. Apart from predicting risk with the aggregated features at a given grade (features of grades 5 and below), we also predicted the risk using individual grade features (features of grade 5). Our results showed that grade 7 features are most relevant, whereas the relevance of the earlier grades sequentially follow.

Second, we showed comparisons between several groups of features by building prediction models individually on these groups. While score features gave the best ROC AUC, demographic features were also found to be important. Yet, more in-depth examination of behavioral features could prove a useful avenue for further research. Our results also showed that while the Math features are most predictive of the risk in grade 8 Math, other subjects are also important predictors.

Third, the solid performance of the logistic regression model, when evaluated against other models, may be relevant as the model is more amenable to coefficient-by-coefficient understanding of variable influences. Additionally, the selection of classifier threshold in a realistic setting is important. Our results showed that a good balance between the true positive and false positive rates can be achieved.

In sum, the breadth of features, extent of data, and approach to modeling make this effort unique and potentially very informative for educational research and practice.

7. FUTURE WORK

We plan to continue our this work towards two fronts: 1) to improve the overall prediction accuracy, and 2) to perform more analysis aimed at deriving further valuable insights for educational institutions.

As described in earlier sections, the data used for prediction contain several missing values. Currently we use mean imputation to approximate the true values of those missing entries. This is a crude approximation and hence is likely to result in noisy predictions. More sophisticated methods to impute missing values are likely to improve the prediction accuracy further.

Currently we are building only one classifier per prediction task for the entire student data. However a divide-and-conquer approach of grouping students into clusters and building a prediction model for each cluster is likely to improve the accuracy further. An interesting aspect to explore in this direction is to group the students' data based on a subset of features that yield optimal risk prediction performance.

Apart from the feature analysis reported in the earlier section, educational institutions would be interested in an explanation of how various predictions were arrived at for each student. This requirement in our system could simply translate into an assignment of importance weights to various features in terms of their role in prediction outcome. Towards this goal, we plan to investigate a detailed feature level discriminant analysis. In addition, exploring alternate models such as hierarchical prediction models might provide an opportunity to back-trace local decisions made at various levels based on local features, thereby making it possible to derive an explanation of why a particular decision is made.

8. CONCLUSIONS

In this paper, we have reported on a large-scale study to predict students at risk of not meeting acceptable levels of performance in national and state-level Grade 8 standardized tests in Mathematics and Science. This study involved one of the largest school districts in the US (GCPS). Using a rich set of predictors related to student demographics and behavior, as well as longitudinal data on test scores through the different grades, we constructed risk prediction models that are able to identify students at risk with a high degree of accuracy. Through experimental evaluation, we also showed that the models strike a good balance between identifying and over identifying students at risk. Other key observations from our experiments include: Predictions for Grade 8 performance may be made with reasonable confidence as early as Grade 4, with data from subsequent grades further improving the accuracy. Amongst the features, longitudinal data on test scores are highly predictive of future success/failure, while a set of demographic features such as ethnicity, disability and free meal, when taken together, are able to achieve reasonably high prediction accuracies as well. Behavioral data do not come up as a strong predictor for performance in this study, and further work is needed to investigate the way behavioral data are collected and interpreted to understand if and why this is actually the case.

Overall, our study shows that longitudinal data-driven risk models for academic performance in standardized tests can lead to robust and early predictions, thereby making it possible to initiate targeted personalized intervention plans well in advance to mitigate the risks and shift a student's learning trajectory towards desired outcomes.

9. REFERENCES

[1] Gwinnet County Public Schools. http://publish.gwinnett.k12.ga.us/gcps/home/public/about.

[2] IBM SPSS Modeler. http://www-01.ibm.com/software/analytics/spss/products/modeler/.

[3] No Child Left Behind. http://www2.ed.gov/nclb/landing.jhtml.

[4] Race to the Top. http://www2.ed.gov/programs/racetothetop/index.html.

[5] Signals Project. http://www.itap.purdue.edu/studio/signals/.

[6] A. J. Bowers, R. Sprott, and S. A. Taff. Do we know who will drop out? a review of the predictors of dropping out of high school: Precision, sensitivity, and specificity. *High School Journal*, 96(2), 2012.

[7] A. B. Brewster and G. L. Bowen. Teacher support and the school engagement of latino middle and high school students at risk of school failure. *Child and Adolescent Social Work Journal*, 21(1):47–67, 2004-02-01T00:00:00.

[8] D. Chen and G. Elliott. Determining key (predictor) course modules for early identification of students at-risk. In *2013 International Conference on Advanced*

Information Engineering and Education Science (ICAIEES 2013). Atlantis Press, 2013.

[9] C. ERDOS, F. GENESEE, R. SAVAGE, and C. HAIGH. Predicting risk for oral and written language learning difficulties in students educated in a second language. *Applied Psycholinguistics*, 35:371–398, 3 2014.

[10] A. Essa and H. Ayad. Student success system: risk analytics and data visualization using ensembles of predictive models. In S. Dawson, C. Haythornthwaite, S. B. Shum, D. Gasevic, and R. Ferguson, editors, *LAK*, pages 158–161. ACM, 2012.

[11] A. for Excellent Education. The High Cost of High School Dropouts: What the Nation Pays for Inadequate High Schools. http://all4ed.org/wp-content/uploads/2013/06/HighCost.pdf, Issue Brief November 2011.

[12] L. Fortin, D. Marcotte, P. Potvin, É. Royer, and J. Joly. Typology of students at risk of dropping out of school: Description by personal, family and school factors. *European Journal of Psychology of education*, 21(4):363–383, 2006.

[13] G. Gray, C. McGuinness, and P. Owende. An investigation of psychometric measures for modelling academic performance in tertiary education.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, 2009.

[15] A. Hershkovitz, S. M. Gowda, R. S. Baker, and A. T. Corbett. Predicting future learning better using quantitative analysis of moment-by-moment learning.

[16] J. M. Hintze and B. Silberglitt. A longitudinal examination of the diagnostic accuracy and predictive validity of r-cbm and high-stakes testing. *School Psychology Review*, 34(3), 2005.

[17] L. Macfadyen and S. Dawson. Mining lms data to develop an "early warning system" for educators: A proof of concept. *Computers & Education*, 54(2):588–599, 2010.

[18] R. Rumberger and S. A. Lim. Why students drop out of school: A review of 25 years of research, 2008.

[19] J. Spring. *American Education*. McGraw-Hill Higher Education, 2007.

[20] K. R. Stevenson. Educational trends shaping school planning and design: 2007. *National clearinghouse for educational facilities*, 2006.

[21] A. Vihavainen, M. Luukkainen, and J. Kurhila. Using students' programming behavior to predict success in an introductory mathematics course. In *Proceedings of The Fourth International Conference on Educational Data Mining 2011*, 2011.