Predicting Student Risks Through Longitudinal Analysis

Ashay Tamhane, IBM Research, Bangalore, India
Shajith Ikbal, IBM Research, Bangalore, India
Bikram Sengupta, IBM Research, Bangalore, India
Mayuri Duggirala, Tata Research, Pune, India
James Appleton, Gwinnett County Public Schools, GA, USA

Presentation @ ACM SIGKDD 2014, NYC, NY, USA
26 August 2014

Problem & Motivation

- Education domain is witnessing unprecedented transformation
- K-12 schooling – crucial period in everyone's education life
- One of the major problems at K-12 level – drop-outs
- Poor academic performance – One of the key indicators of drop-out
- Predict potential risks in academic performance for early intervention

Predicting potential risks in performance of the students ahead in time!

Predicting Potential Risks in Academic Performance

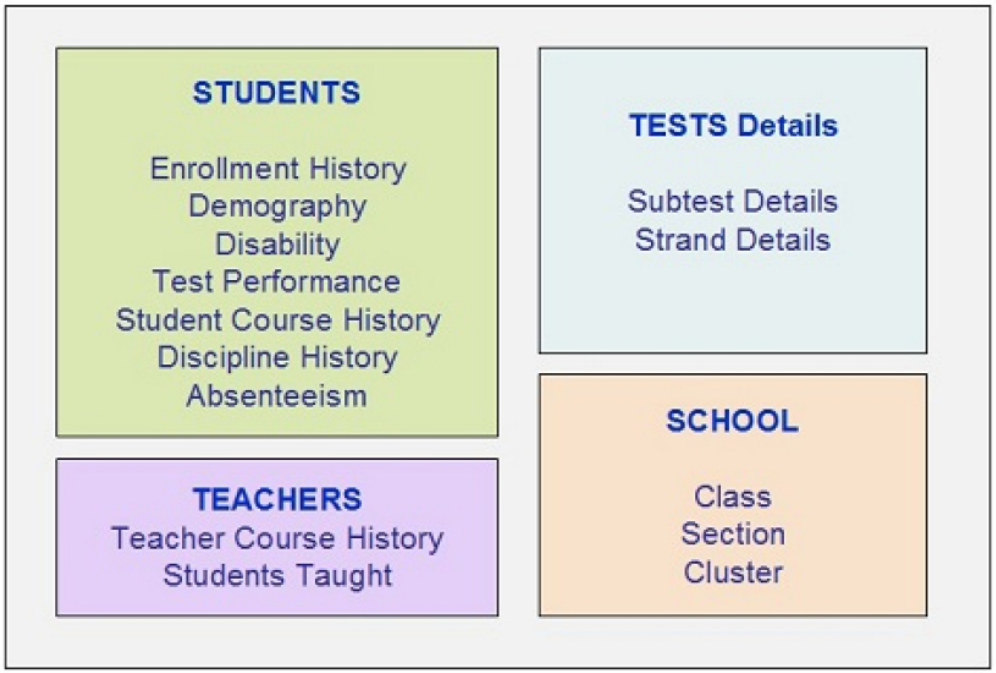
- Traditionally – Teachers predict
  - Using recent past academic results, experience with similar students in the past
  - Negatives:
    - limited knowledge, not objective quantification
    - Often do not leave enough time to apply appropriate intervention
- Now – There is an opportunity to predict better and well ahead in time
  - With digitization of school records and the use of instrumented digital learning environments
  - Student's longitudinal journey through K-12 is captured
  - Data from thousands of students from the past is available
    - Including academic history and non-academic attributes such as demography, behavior.

This is what we tried to do in this work!!

• In collaboration Gwinnett County Public Schools

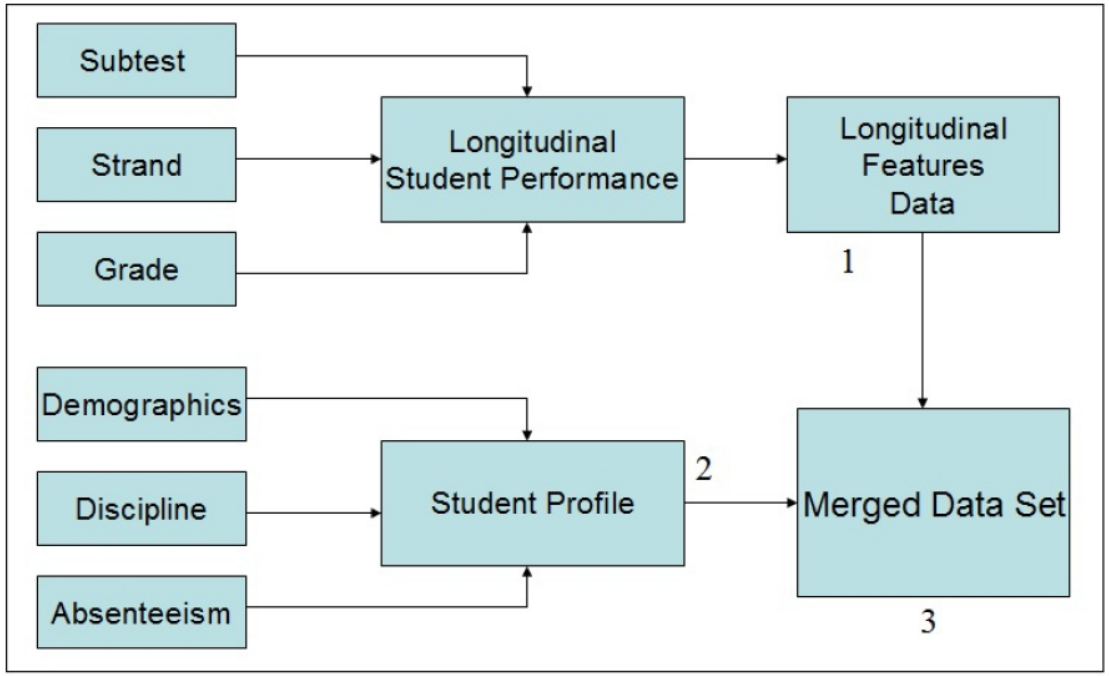
Data from Gwinnett County Public Schools

- One of the largest school districts in the US
  - 132 schools, serving ~168,000 students per year.
  - Data related to students, teachers and assessments from all constituent schools are collated into hundreds of tables in a central data warehouse.
  - A snapshot of this warehouse was made available to IBM.

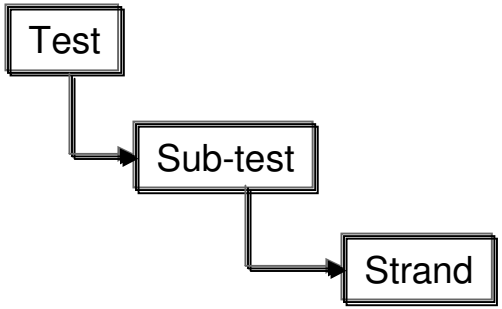


Specific Data Considered

- Grades: 1 to 8 (Primary & Middle school)
- Subjects: Mathematics, Science, Literature,
- Tests:
  - CRCT – Criterion References Competency Test
  - ITBS – Iowa Test of Basic Skills
  - CogAT – Cognitive Ability Test



Test Hierarchy



Longitudinal view includes:
- scores from all past grades, tests, subtests, and strands

~ 160,000 students max. 516 scores per student

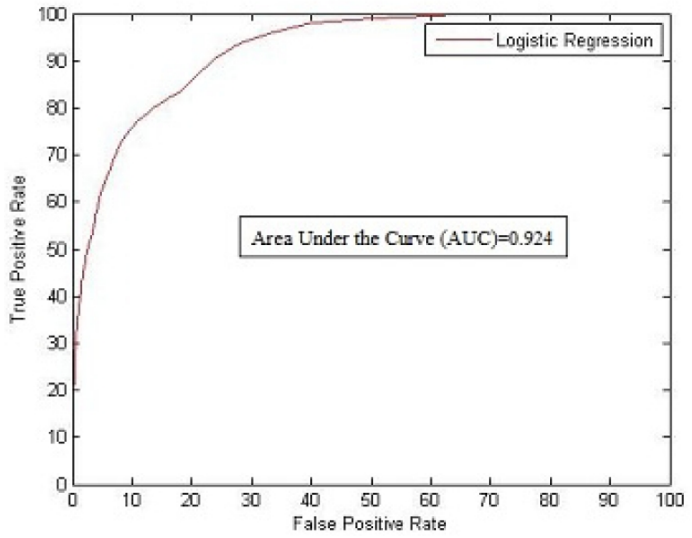
Many missing scores!!

Prediction Task

- Targets considered:
  - CRCT 8th Grade Mathematics
  - CRCT 8th Grade Science
  - ITBS 8th Grade Mathematics
- Data Preparation:
  - Target: for CRCT score < 800 is considered 'at-risk'. For ITBS score < 25 is 'at-risk'
  - Features: all scores from grades < 8th grade + demography + behavior – many scores missing
  - Students chosen such that at least 20% features are present
  - Missing features are mean imputed
  - Data size: CRCT - 58707 students and 342 features; ITBS - 43310 students and 282 features
  - Experimental setup: 5-fold cross validation
- Prediction:
  - Classifiers from IBM SPSS or WEKA: logistic regression, naïve bayes, decision tree
  - To predict: 'at-risk' and 'no-risk' students.
- Evaluation metric:
  - ROC-AUC – area under receiver operating curve - true positives vs false positive
  - False positive rate for True positive rate of 90% or more

Risk Prediction Performance

Sample ROC curve →



ROC-AUC for various classifiers

↓

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

FP for TP>=90 →

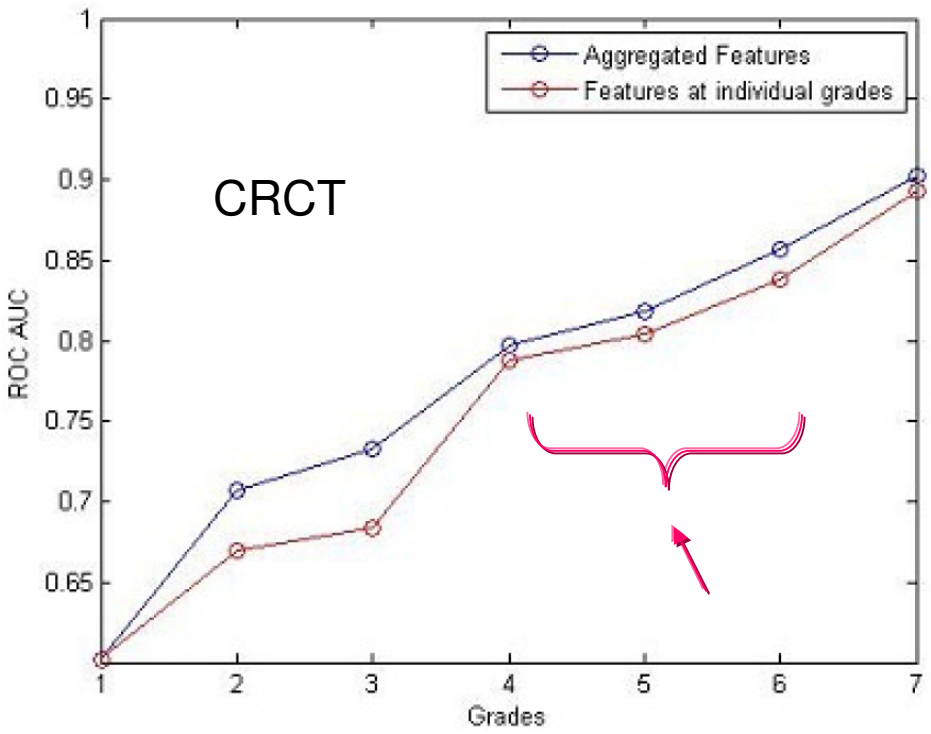
Task	Probability Threshold	True Positive TP, in %	False Positive FP, in %
CRCT 8th Mathematics	0.06	90.5	23.8
CRCT 8th Science	0.18	90.0	24.7
ITBS 8th Mathematics	0.1	90.7	28.8

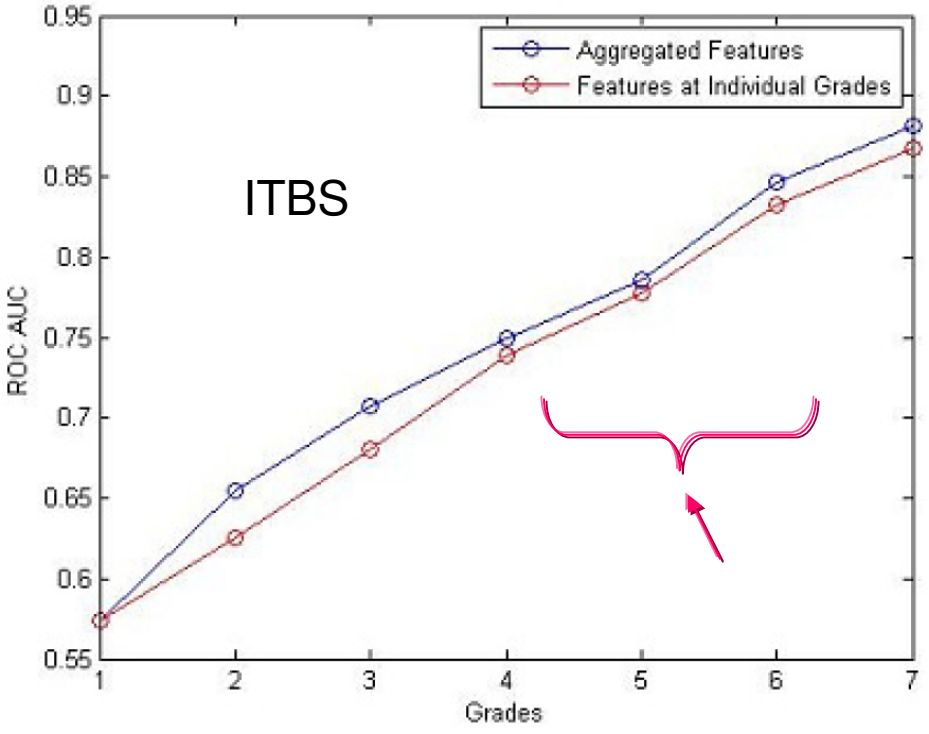
Feature Importance

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

- **Scores are important, demography information helps**
- **Recent past scores are the most important**

Early Prediction





- At Grade 4, it is possible to predict for Grade 8 with reasonably high accuracy
- Accuracy improves as more and more features are aggregated from lower grades

Summary

- Problem: Predicting students at risk of poor academic performance
  - To facilitate planning of effective personalized interventions
- Conclusions from our study
  - It is possible to predict at-risk students with high accuracy
  - Past scores are important indicators – recent past scores are more important
  - It is possible to predict well ahead in time – thus providing enough time for effective interventions.
- Highlight of our work
  - The scale of our study, large amount of data from major US school district (Gwinnett County)
- Potential future directions
  - To expand this to other grades / subjects – taking in all other features available
  - Prediction accuracy improvement
    - Improve missing value handling
    - Estimate student clusters and build prediction model per cluster
  - Feature importance – reasoning out a prediction
    - Discriminant analysis
    - Hierarchical prediction models to back-trace local decisions

Thank you!

