UTSA
The University of Texas at San Antonio™

A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes

Presented by: Amjad Alqahtani

Interventions to address adverse academic outcomes

- Goal:
  - Graduating high school on time
- Challenges:
  - Lack of skills, motivations, health, economical, personal, and various other factors
- Solution:
  - Provide appropriate interventions

How do we identify students in need of interventions?

Traditional way of Identify Students At-Risk

- Schools relied on feedback from instructors
- GPAs
- Absence rates
- tardiness
- Cansolar discuss with students
- Report Cards
- etc

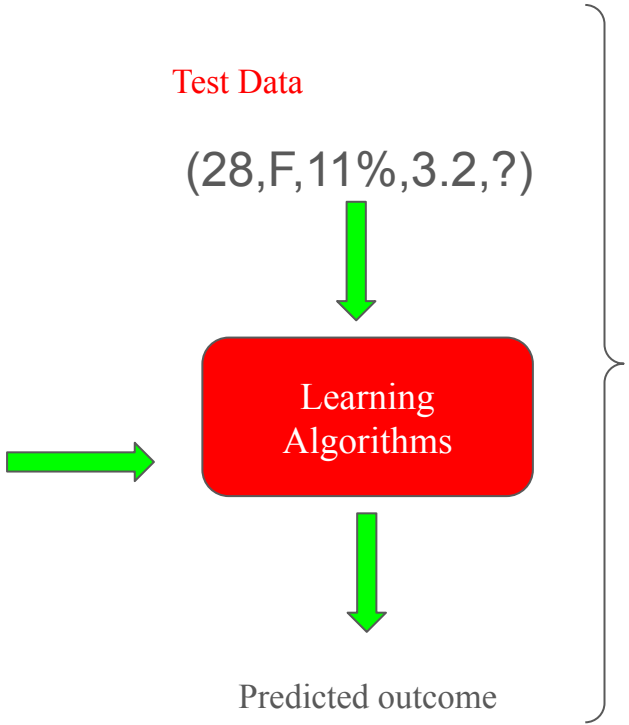
All these methods(involved human judgments) for assessing risk are expensive, time consuming and error prone.

Alternative way of Identify Students At-Risk

- Rule based System using ML such as:
  - Logistic Regression
  - Decision Trees
  - Random Forests.

Age	Gender	Absence rate	GPA	No-grad
15	F	8%	3.8	0
14	F	9%	2.6	0
16	M	19%	2.8	1

Outcome variable



Evaluating Learning Algorithm

- AUC of the ROC
- Precision
- Recall
- Accuracy

School Settings

- Resources for intervention are limited and vary highly with time.

- Predicting risk as early as possible.

- School administrators want to know what kind of risk and mistake that learning algorithm have

Dataset

- Datasets from two different school districts

Districts	Students	Grades	Attributes
District A	150,000+(mid-atlantic region, 40 schools)	6th - 12th	Gender, Age, Ethnicity, City, School Absence Rate, Cumulative GPA, English Proficiency, Disability Status, Retained Status, Did not Graduate on time?
District B	30,000+(east coast,39)	8th - 12th	

- Multiple cohorts of students used as train and test sets

Approach 1: Risk Score: Predicting At-Risk Students

- A probabilistic estimate indicating the likelihood that a student will not graduate on time, generated by machine learning models.
- Models like Random Forest, Logistic Regression, and others calculate risk scores based on students attributes like GPA
- Two metrics:
  - Empirical Risk Curves: Evaluates how well the risk scores rank students.
  - Precision and Recall at top k: Measures accuracy when selecting the top at-risk students for intervention.

Approach 1: Risk Score: Predicting At-Risk Students

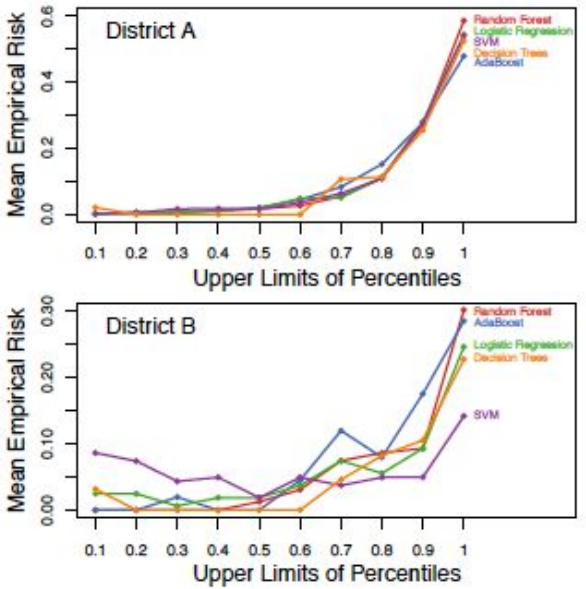


Figure 2: Empirical Risk Curves. The ranking quality of an algorithm is good if this curve is monotonically non-decreasing.

Approach 2: Evaluation of Timely Risk Detection

- Identifying Risk Early:
  - Detecting student risk by the end of middle school is preferred over detecting it during the final year of high school.
  - Plot Precision at 5%: Use data collected up until the end of various grades to plot Precision at Top 5%, helping track the accuracy of early predictions.
- Identifying Risk Before Students Go Off-Track:
  - Detecting risk before a student starts failing grades allows for more effective interventions.
  - Key Metric:
    - Ratio: Number of at-risk students identified before failing grades, compared to the total number of students who did not graduate on time.

Approach 2: Evaluation of Timely Risk Detection

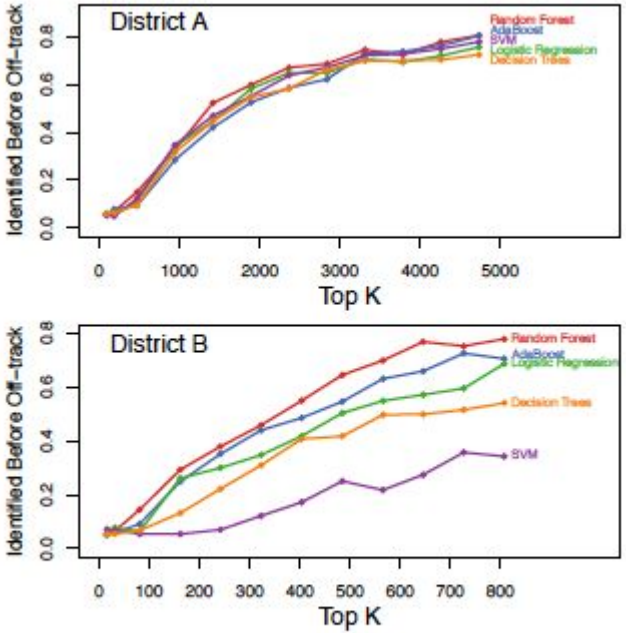


Figure 8: Identification before Off-track

Approach 3: Characterizing Patterns of Mistakes

- Identify **frequent itemsets** of triples (attribute, relation, value) using the **FP-Growth algorithm**.
- Rank students based on their **risk scores** and set the predicted value of **no_grad** (not graduating on time) to **1** for the top **K** high-risk students.
- Create a new field called **mistake**. Set this field to **1** if the predicted value of **no_grad** does not match the actual value (i.e., the prediction was wrong).
- For each frequent pattern, compute the **probability of a mistake**.
- Rank the patterns based on these **mistake probabilities**, prioritizing those with the highest likelihood of error.

Qualitative Analysis of Mistake Pattern

Pattern mistake	Description
High GPA but High Absence Rate	• Random Forest misclassified students with GPA > 3.0 and absence rate > 30% as low-risk.
Disability, Female, GPA ≥ 3.5	Decision Tree misclassified female students with a disability and GPA ≥ 3.5 as low-risk.

4. Features used for Risk Assessment

- Logistic Regression:
  - Features are ranked based on the magnitude of the coefficients.
  - Key Feature: Gender (Male) and Absence Rate (Grade 8) are heavily used in predicting risk.
- Decision Trees:
  - Features are ranked based on Information Gain.
  - Key Features:
    - GPA (Grade 8)
    - GPA (Grade 7).
- Random Forest:
  - Features are ranked based on Information Gain, averaged over all the trees in the ensemble.
  - Key Features:
    - GPA (Grade 7)
    - Absence Rate (Grade 7).

Conclusions

- An exhaustive analysis bridges the gaps between evaluations meaningful to educators and traditional machine learning metrics.

- Outlined several metrics catering to various aspects such as reliable risk estimation, early prediction, and interpretability.

- Evaluation framework has been very useful in deploying machine learning solutions for both school districts