Identifying Resilience Factors in Texas Public Schools

http://DataLab12.github.io/

June Yu, Daniel Payan, Dr. Jelena Tešić

Department of Computer Science

Texas State University

Dr. Li Feng

Department of Finance and Economics

Motivation

- COVID-19 school reopening decisions were difficult for policymakers since there was no consensus on the impact of school reopening on the spread of COVID-19
- Learning loss was documented in many states including Texas
- **If we can identify most impactful factors on learning loss from publicly available data sources during pandemic, we can help policy makers make more informative decisions on learning recovery**

Research Questions

- Can we quantify the impact of the mode of instruction(hybrid, remote, in-person) on the learning loss?
- Do school district reopening decision influence the learning loss experienced by students?
- Are students from low-income background and minority students experience more learning loss?
- Do students from different grade level experienced learning loss differently?

Data Acquisition and Integrations

Data are acquired from 7 different sources below and integrated by matching School District ID and County FIPS Code with 79 variables from 1,165 school districts in 253 counties:

- STAAR test results, math and reading, by grade in 2019 and 2021 from the Texas Education Agency
- COVID case data, # of students on campus reported to the Texas Health and Human Services per county
- Student race/ethnicity, Title 1/Free lunch, Teacher-Student ratio per district from Common Core Data from the National Center for Education Statistics(NCES)
- Local Area Unemployment Statistics(LAUS) per county from U.S. Bureau of Labor Statistics
- Average Daily Attendance(ADA) per district from Texas Education Agency
- 2010 Census Block Group data from Texas Education Agency/Census Bureau
- Elementary and Secondary School Emergency Relief(ESSER) Grant from Texas Education Agency

Exploratory Data Analysis

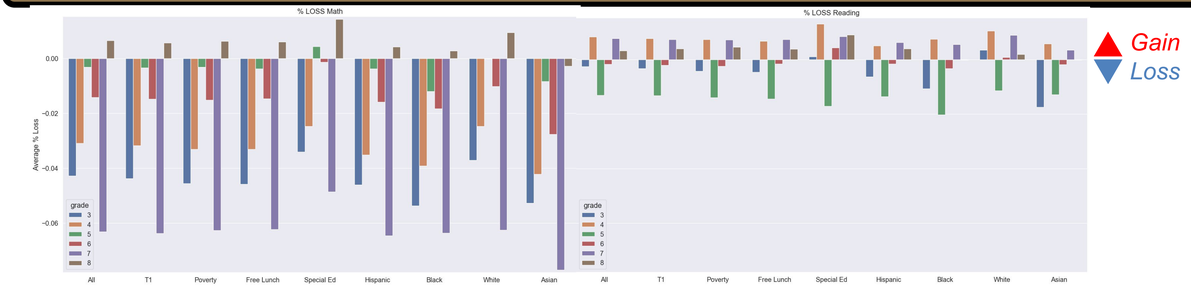

Figure 1: Learning Loss % for Math(left) and Reading(right) for group of students: Title 1, Poverty, Free Lunch, Special Ed, Hispanic, Black, White, Asian

- Learning loss is calculated by getting STARR score differences $\frac{Average\ Score\ 2021 - Average\ Score\ 2019}{Average\ Score\ 2019}$
- Math shows more severe learning loss throughout the most student groups compared to Reading
- 3 classes label has been created:

Loss < 25th	25th percentile ≤ Expected ≤ 75th percentile	75th < Gain

Impactful Factors

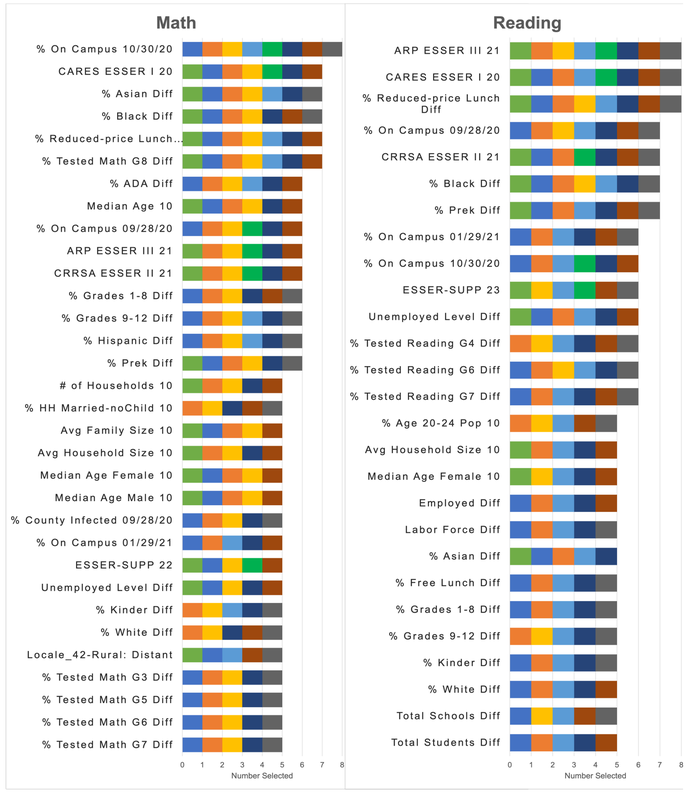

Figure 2: Number of Predictors Selected by 9 feature selection methods

The most impactful predictors are identified using 9 different feature selection methods:

- Filter Methods:
  - Variance Threshold
- Embedded Methods:
  - L1 (Lasso) Regularization
  - Random Forest Feature Importance
- Wrapper Methods:
  - Permutation Importance - Random Forest
  - Permutation Importance - Ridge
  - Recursive Feature Elimination - Random Forest
  - Recursive Feature Elimination - Ridge
  - Sequential Feature Selection - KNN
  - Sequential Feature Selection - Ridge

Findings:

- The most impactful predictors for math are. **the ratio of students on campus on 10/30/20** Covid aid in 2020, student's race, reduced-price lunch eligibility
- **The most impactful predictors for reading are Covid aid given in 2020 and 2021**, reduced-price lunch eligibility, and student's race, the student ratio on campus on 09/28/20 and the ratio of pre-k students.

Gradient Boosting

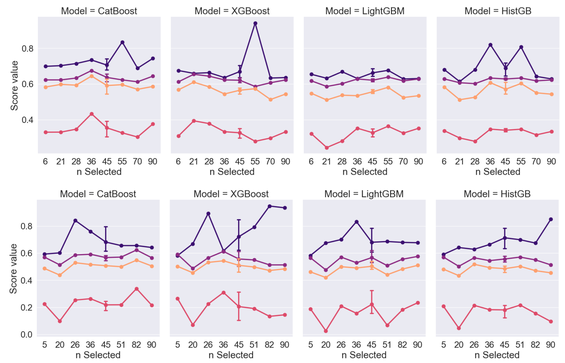

Figure 3: Four Gradient Boosting Models scores for Math (top) and Reading (bottom)

Train Accuracy Test Accuracy MCC F1

- The state-of-an-art gradient boosting models were built on the feature sets selected from 9 methods to examine dimensionality reduction effects on predicting learning loss
- Hyperparameters such as tree structure constraints, learning rate, L2 regularization were tuned to avoid overfitting and increase accuracy
- Feature space does not have significant influence for performance of the gradient boosting models

Conclusion

- Add STARR exam scores for 2022 to confirm the resilience factors effects
- Update Census Block Group data for 2020 to grasp the characteristics of socioeconomic factors up-to-date
- Compare outcome for missing values and pre-processing approaches

Acknowledgements

The work has been supported by Community Health and Economic Resilience Research (CHERR) @ Data Lab (DataLab12.github.io)