A Data-Driven Approach to Understanding and Predicting Chronic Absenteeism in K–12 Education

Chase Clemence, David Corcoran, Gentry Lamb, Adam Stein
Georgetown University - Data Science and Analytics

Introduction

Chronic absenteeism in school districts across the U.S. is a growing concern, as it affects student performance and success later in life. A student is considered chronically absent if they miss at least 10% of school days. The COVID-19 pandemic disrupted education in the U.S. and has caused higher rates of chronic absenteeism, even as schools have reopened. This study aims to predict which school districts are prone to chronic absenteeism by leveraging demographic and financial data from the 2022–2023 school year.

Goals:

- **Classify** school districts as having low or high chronic absenteeism using demographic, financial, and school support-related features
- **Identify key predictors** of chronic absenteeism across U.S. school districts
- **Inform policy** by translating model insights into actionable recommendations for reducing chronic absenteeism

Methods and Materials

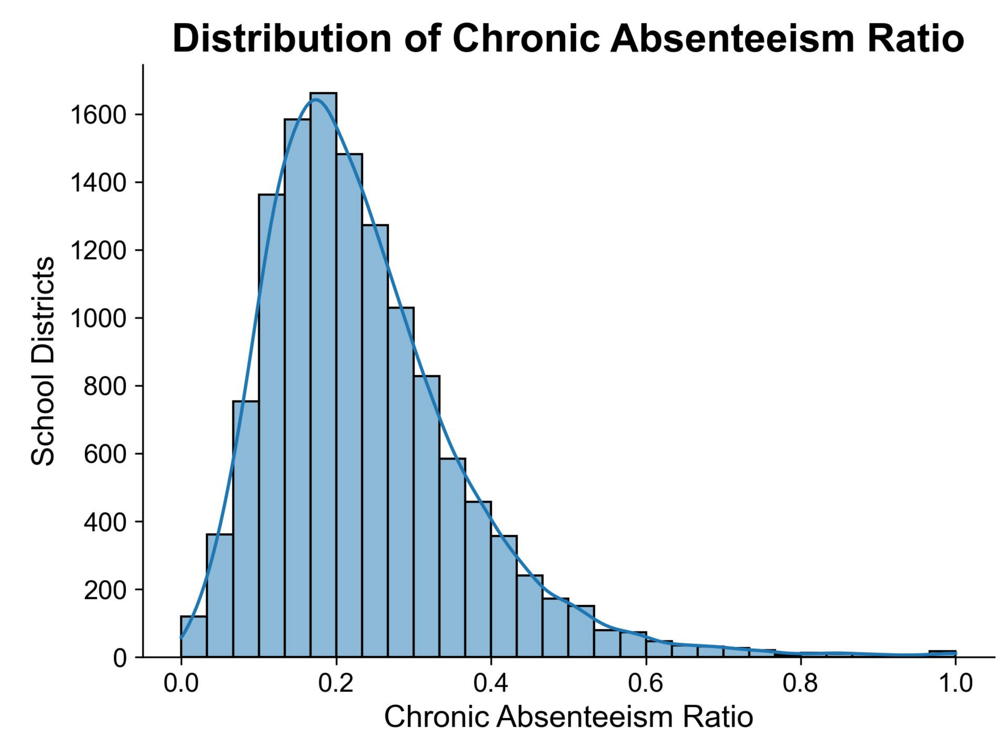
This study used publicly available data from the 2022–2023 school year, compiled from several sources. **Total size ~12,800 rows.**

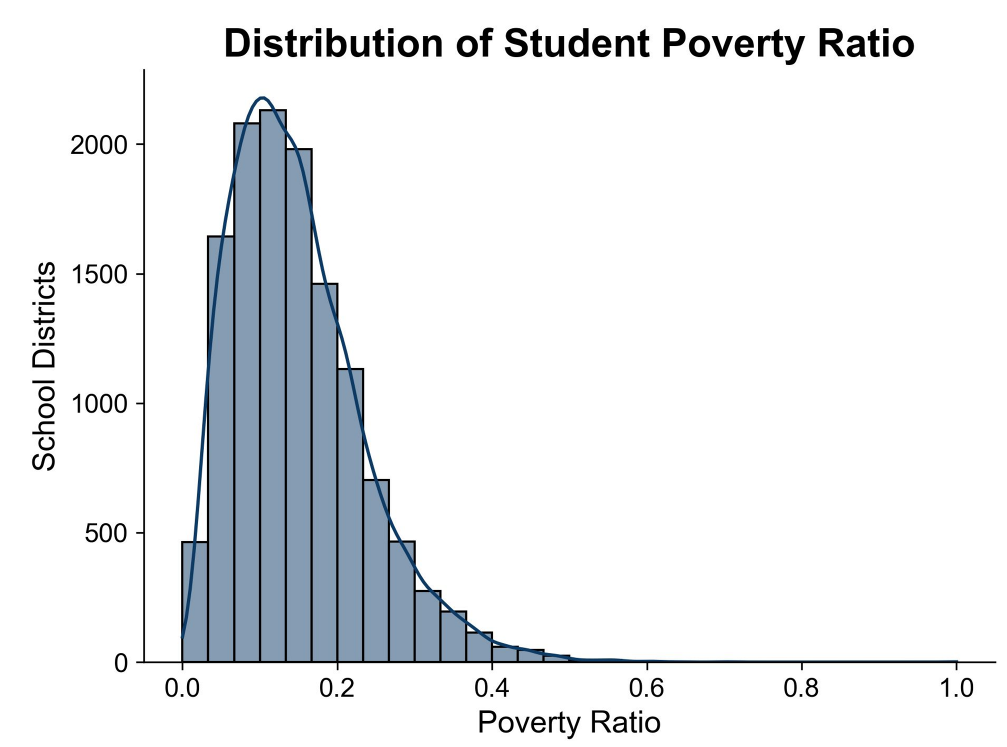
- **U.S. Department of Education** – Chronic absenteeism rates by district
- **U.S. Census Bureau (SAIPE)** – District-level poverty estimates for K–12 populations
- **NCES** – Racial demographics and financial data from the Common Core of Data

Classification models used in this study:

- **Logistic Regression** – Models absenteeism probability by linearly combining features
- **Support Vector Machine** – Finds the best hyperplane to separate low vs. high absenteeism school districts
- **LDA/QDA** – Classifies based on group distributions; QDA allows more flexibility
- **Random Forest** – Combines decision trees and ranks feature importance
- **Neural Network** – Captures complex, non-linear patterns in the data

Data Exploration



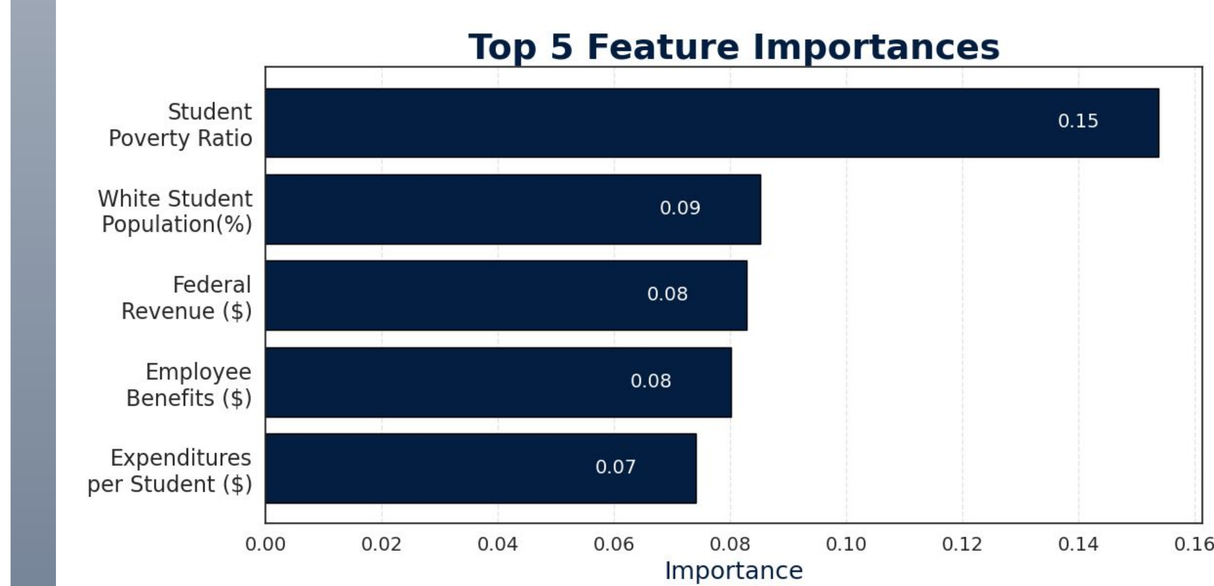


Modeling Summaries

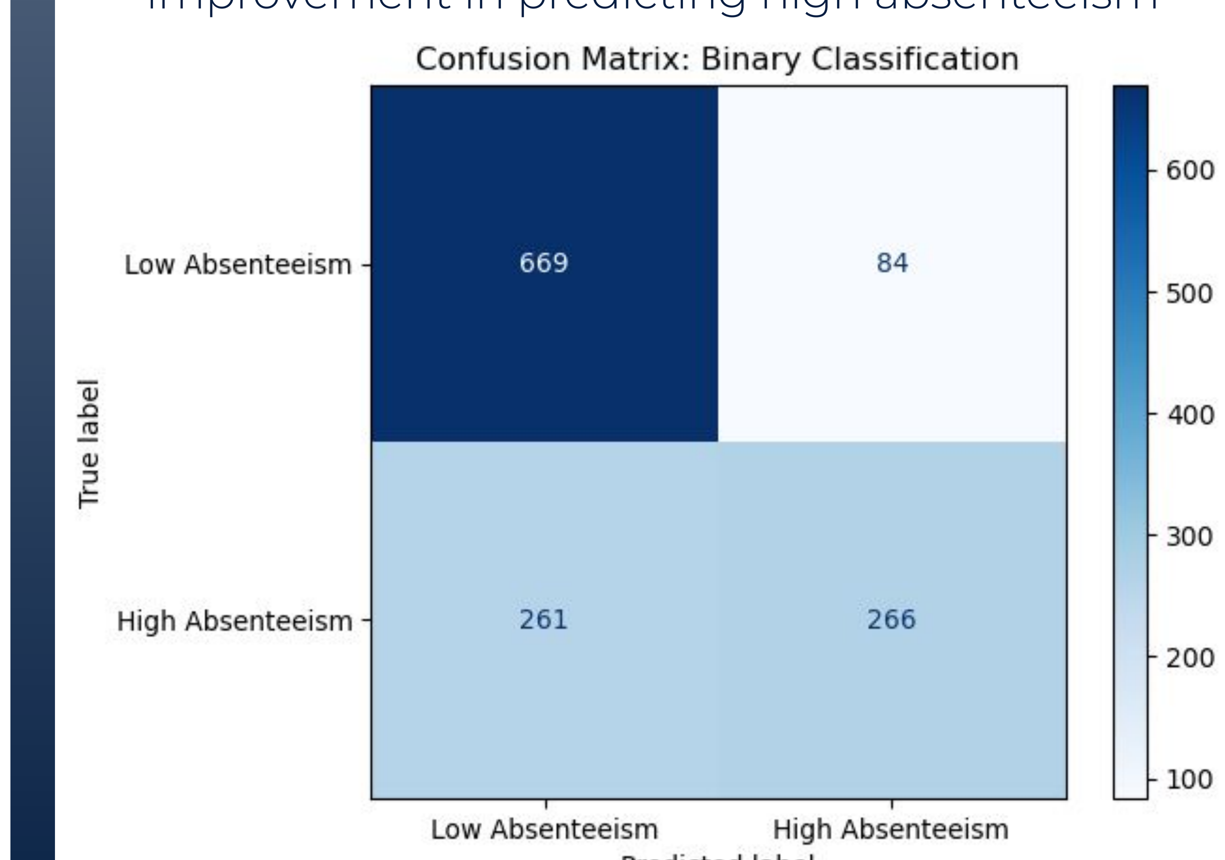
Model	Accuracy	AUC
Logistic Regression	74%	0.711
SVM	65%	0.664
LDA	67%	0.756
QDA	64%	0.730
Random Forest	74%	0.728
Neural Network	73%	0.815

Results

- **Neural Network:** Best model based on Accuracy and AUC; more complex to tune/interpret
- **Random Forest:** Next best model; helped determine and rank variable importance
- **Logistic Regression:** Good performance with high interpretability
- **SVM & LDA/QDA:** Lower performance, struggled with imbalanced classes



- **AUC:** Better than accuracy for imbalanced data; shows how well the model distinguishes classes
- **Feature Engineering:** Log transformations and scaling boosted performance, especially for linear models
- **Hyperparameter Tuning:** Improved Neural Network performance by optimizing layers, learning rate, and batch size
- **Confusion Matrix (NN):** Highlights the trade-off between precision and recall; room for improvement in predicting high absenteeism



Discussion

Limitations:

- **Class Imbalance** - Fewer positive cases made high absenteeism hard to predict
- **Feature Relationships** - High feature overlap limited interaction modeling
- **Generalizability** - Local differences may limit broader application

Takeaways:

- **Poverty** and **race demographics** are consistent predictors of absenteeism
- Ensemble methods and neural networks **outperform linear models** in accuracy but sacrifice interpretability

Future Work:

- Address **class imbalance** with SMOTE or re-weighting
- Explore **time trends** (e.g., seasonal absenteeism)
- Improve **feature engineering** and test region-specific models

Conclusion

- **Poverty emerged as the most consistent predictor** of high absenteeism, bolstering the connection between economic hardship and school attendance
- **Funding and support services** should be allocated to schools where **poverty-related absenteeism** is most prevalent
- Model predictions should be leveraged to **identify at-risk students early** and implement proactive outreach programs

References

[1] Chronic absenteeism. U.S. Department of Education. (2025, January 20). https://www.ed.gov/teaching-and-administration/supporting-students/chronic-absenteeism
[2] Elsi - Elementary and secondary information system. (n.d.). https://nces.ed.gov/ccd/elsi/
[3] Bureau, U. C. (2022, July 1). Small area income and Poverty Estimates (SAIPE) program. Census.gov. https://www.census.gov/programs-surveys/saipe.html