# ASDS 6302

# Final Project Report

# TEXAS SCHOOL RATING ANALYSIS

# Texas Flag Svg, Texas Flag Png, Texas Flag Cut File, Texas Layered Flag Svg - Etsy

# Group 2

**Team Members:**

Lingala Harini Sarma – [hxl5602@mavs.uta.edu](mailto:hxl5602@mavs.uta.edu)

# Ahmed, Mohamed - [mohamed.ahmed2@mavs.uta.edu](mailto:mohamed.ahmed2@mavs.uta.edu) Kondabathini, Sai Harshith - [sxk7657@mavs.uta.edu](mailto:sxk7657@mavs.uta.edu) Trinh, Phuong Le Thien - [phuong.trinh2@mavs.uta.edu](mailto:phuong.trinh2@mavs.uta.edu)

## Introduction

## This project examines public school performance across Texas using official data from the Texas Education Agency (TEA). Our primary objective was to predict the Overall Rating (A–F) of schools based on academic, demographic, regional, distinction-based, and engineered features. The project integrates Python-based modeling and Power BI visual analytics to provide both statistical and practical insights into school performance across Texas.

## Dataset Source

The dataset used for this project was obtained from the **Texas Open Data Portal**, provided by the **Texas Education Agency** **(TEA)**. It contains official statewide accountability ratings for the 2022–2023 school year.

**Dataset Title:** School Year 2022–2023 Statewide Accountability Ratings

**Source:** Texas Open Data Portal (data.texas.gov)

**Link:** <https://data.texas.gov/dataset/School-Year-2022-2023-Statewide-Accountability-Rat/nui6-x374/about_data>

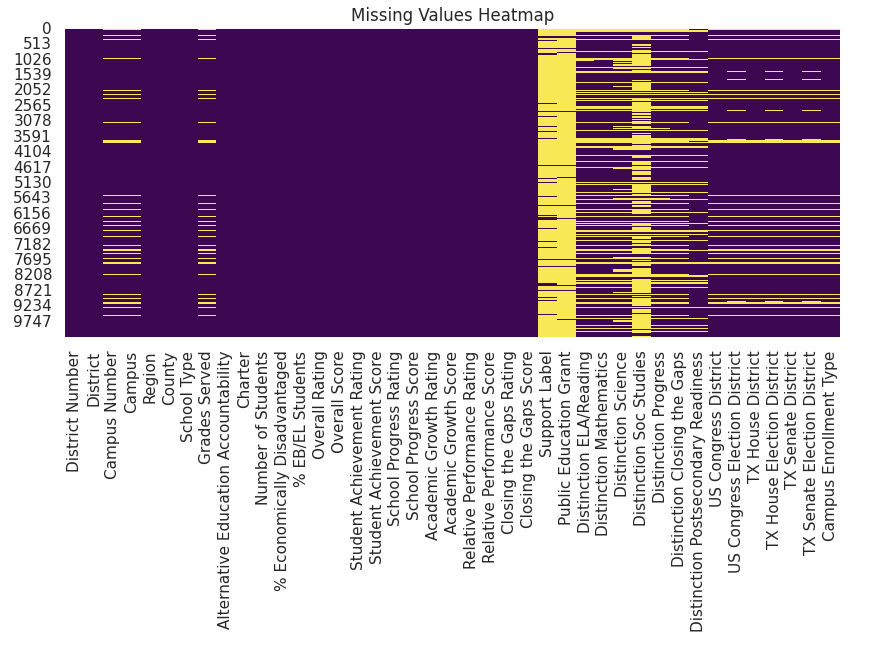
The dataset includes school identifiers, academic indicators, distinctions, demographic information, and the Overall Rating, which served as the target variable for our machine learning models.

## Data Overview

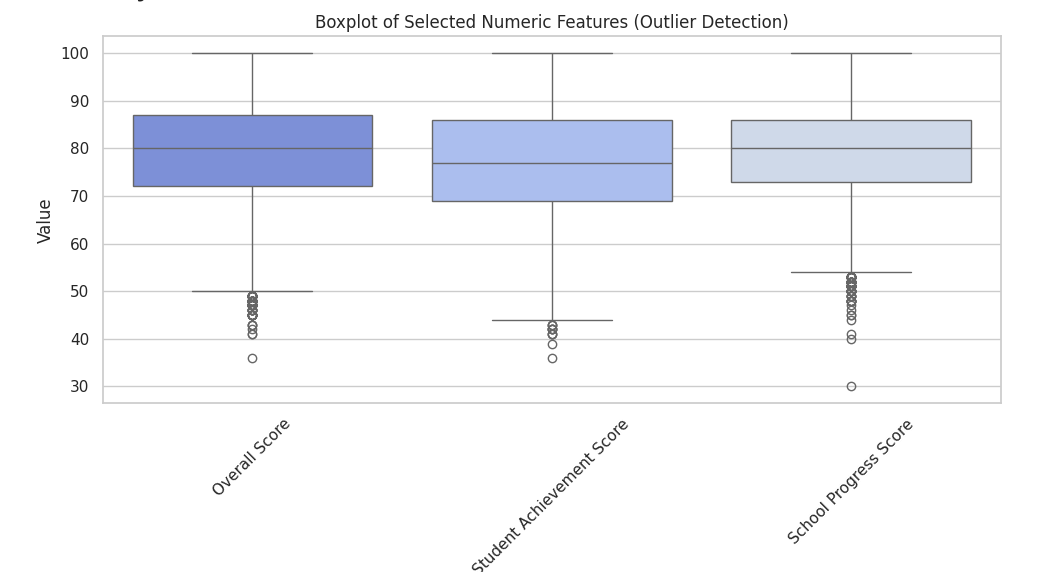
The dataset used in this project contains comprehensive information on more than 10,000 public schools and nearly 1,000 districts across Texas for the 2022–2023 academic year. It includes detailed geographical identifiers such as district, county, and region, as well as school-level characteristics including school type (elementary, middle, or high school) and charter status. In addition, the dataset provides demographic indicators like the percentage of economically disadvantaged students and English learners, alongside key academic performance metrics such as School Progress, Academic Growth, and Closing the Gaps. The TEA distinction designations—covering subjects like mathematics, reading, science, and postsecondary readiness—are also included as important predictors of school quality. The target variable for modeling, the Overall Rating (A–F), was converted to a numeric scale from 4 (A) to 0 (F) to support machine-learning analysis. Overall, the dataset offers a rich and diverse set of features that enable a thorough examination of school performance across Texas.

## Data Preprocessing

The data preprocessing stage focused on preparing the TEA dataset for reliable and accurate modeling. We began by removing columns that contained more than 80% missing values, as these provided limited analytical value.



For the remaining variables, missing data were handled using mean imputation for numeric fields and mode imputation for categorical fields, resulting in a fully complete dataset with zero missing entries. A duplicate check confirmed that no repeated rows were present. Although outliers were detected using the IQR method, they were retained because they represent real variations in school performance rather than data errors.



Next, categorical variables such as district, region, and school type were encoded appropriately, and the target variable, Overall Rating (A–F), was converted into a numerical scale ranging from 4 to 0. Finally, continuous features were standardized using the StandardScaler to support models that benefit from normalized data, such as Logistic Regression and LDA. These preprocessing steps ensured that the dataset was clean, consistent, and ready for effective machine-learning analysis.

## Exploratory Data Analysis (EDA):

To better understand school performance across Texas, we performed an extensive Exploratory Data Analysis (EDA) using both Python summaries and a multi-page interactive Power BI dashboard. The goal of this EDA was not only to visualize patterns in school ratings but also to uncover deeper geographic and demographic insights by drilling down from statewide trends to specific regions, counties, and districts.

A screenshot of a computer

AI-generated content may be incorrect.

We began at the highest level by examining the statewide distribution of school ratings. The bar chart shows that B-rated schools dominate Texas, followed by C-rated and A-rated schools. D and F ratings appear much less frequently, indicating that most Texas campuses are performing at or above average. When breaking this down by school level, elementary schools display the strongest performance, with the highest number of A ratings statewide, while high schools and middle schools show more mixed results. The comparison between charter and non-charter schools highlights distinct patterns as well—charter schools exhibit more variability, showing both high-achieving and lower-performing campuses.

We also analyzed TEA distinction categories using the donut chart, which revealed that Postsecondary Readiness, Closing the Gaps, and ELA/Reading distinctions are the most earned across Texas. This suggests that many schools excel in college for readiness and core academics, while distinctions such as Social Studies or Math Science appear slightly less frequent.

After exploring statewide trends, we used Power BI’s interactive map visualizations to dive deeper into geographic patterns.

A map of the state of texas

AI-generated content may be incorrect.

The **Region Map** clearly indicates that clusters of **A-rated schools are highly concentrated around major metropolitan areas**—Dallas–Fort Worth, Houston, Austin, and San Antonio. These regions stand prominently, showing dense groupings of top-performing campuses. This visualization allowed us to quickly identify the strongest regions in Texas.

From the region-level view, we used the drill-through capability to move into specific counties, allowing us to pinpoint where the highest concentrations of A-rated schools are located.

A map of a state

AI-generated content may be incorrect.

One of the most notable clusters appears in the Fort Worth metropolitan area, situated in Tarrant County. The map visualization reveals a dense concentration of high-performing schools in communities such as Fort Worth, Keller, Grapevine, North Richland Hills, and Arlington, immediately positioning Tarrant County as one of the strongest-performing regions in Texas.

A screenshot of a computer

AI-generated content may be incorrect.

Continuing our analysis, we drilled further into the District Page, which offers detailed information about the distribution of A-rated schools across individual school districts within Tarrant County. The map on this page displays sizable clusters of high-performing campuses across districts including Keller ISD, Fort Worth ISD, Arlington ISD, and Grapevine-Colleyville ISD. The bubble map uses bubble size and color to distinguish school levels—elementary, middle, and high school—making it easy to identify where A-rated schools are most concentrated.

Additional visuals on this page reinforce these findings. The bar chart shows that elementary schools make up most A-rated campuses in Tarrant County, with middle and high schools following closely behind. The distinctions bar chart highlights that high-performing schools in this area frequently earn recognitions in Postsecondary Readiness, Closing the Gaps, Reading/ELA, and Mathematics, demonstrating strong academic preparation and equitable student outcomes. Finally, the charter comparison donut chart shows that most A-rated schools in Tarrant County are non-charter campuses, underscoring the strong performance of traditional public-school districts in the region.

1. **Feature Engineering**

To enhance model performance and capture additional dimensions of school quality, we developed several engineered features based on existing academic and demographic indicators. These features were designed to highlight patterns that traditional variables may not fully represent. The Performance Gap Index measures the disparity between achievement outcomes and equity-based metrics such as Closing the Gaps, helping identify schools with inconsistent performance across student groups. The Achievement Efficiency Ratio evaluates how effectively a school performs relative to its size and percentage of economically disadvantaged students, providing insight into whether campuses are “doing more with less.” Finally, the Distinction Rate converts the total number of TEA distinctions earned by a school into a normalized 0–1 scale, offering a clearer representation of how consistently a school excels across academic categories. Together, these engineered features add depth to the dataset and improve the model’s ability to identify high- and low-performing campuses.

## Feature Selection

We selected the final model features using a combination of correlation analysis, interpretability considerations, and domain relevance. Strong academic indicators—such as Student Achievement, Academic Growth, and School Progress—were retained due to their high correlation with the Overall Rating and their critical role in TEA’s accountability system. Demographic factors, including the percentage of economically disadvantaged and English learner students, were kept to account for socioeconomic influences on school outcomes. We also included TEA distinction indicators and the engineered features (Performance Gap, Achievement Efficiency, Distinction Rate), as they consistently demonstrated meaningful relationships with school performance. Features that showed low relevance, high redundancy, or minimal predictive value were removed to reduce noise and improve model efficiency. This systematic approach resulted in a streamlined and well-balanced feature set that supports accurate and interpretable prediction of school ratings.

A screenshot of a graph

AI-generated content may be incorrect.

## Model Development

After completing feature engineering and selection, we proceeded to develop machine-learning models that could accurately predict the Overall Rating (A–F) of Texas public schools. To ensure a fair and consistent evaluation across all algorithms, the dataset was split into 80% training data and 20% testing data. This split was chosen to provide sufficient data for learning complex patterns while maintaining a sizable, unseen test set for unbiased performance assessment.

We implemented four supervised classification models: Logistic Regression, Linear Discriminant Analysis (LDA), Random Forest, and XGBoost. These models were selected to represent a broad methodological spectrum—from linear classifiers to advanced ensemble techniques—allowing us to evaluate how different types of algorithms respond to the academic, demographic, and distinction-based features in the dataset.

The linear models (Logistic Regression and LDA) served as interpretable baselines, offering insights into feature directionality and general class separation. In contrast, the ensemble models (Random Forest and XGBoost) were equipped to capture nonlinear patterns and complex interactions among variables, particularly those introduced by the engineered features such as Performance Gap, Distinction Rate, and Achievement Efficiency. Each model was trained using the standardized feature matrix and evaluated using consistent performance metrics to enable a direct comparison of predictive strength.

## Model Performance

**9.1 Logistic Regression**

Logistic Regression provided a strong baseline for comparison, achieving approximately 65% accuracy with a macro-averaged ROC–AUC of 0.89. The model performed adequately in identifying middle categories such as B and C but struggled with distinguishing extreme classes (A and F). These limitations are expected given the model's linear decision boundaries and the nonlinear structure of school performance data. Nonetheless, Logistic Regression contributed meaningful interpretability and served as a useful reference point.

A graph of numbers and a number

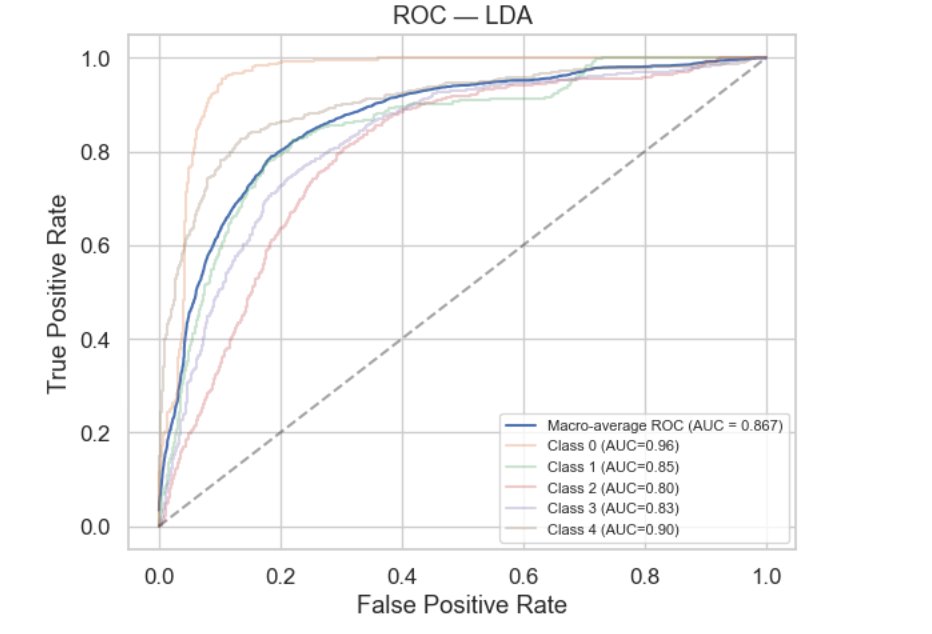
AI-generated content may be incorrect.A graph of a logistic regression

AI-generated content may be incorrect.

**9.2 Linear Discriminant Analysis (LDA)**

LDA achieved an accuracy of around 60% and an ROC–AUC of 0.86. As a probabilistic model, LDA offered better-defined class boundaries than Logistic Regression, but its assumptions of normally distributed features and equal covariance matrices constrained its performance. Misclassification was most prominent among adjacent categories (B, C, and D), reflecting the overlapping academic profiles of schools in the middle performance range. LDA provided valuable structural insight but was less effective for high-precision prediction.

A graph of a number

AI-generated content may be incorrect.

**9.3 Random Forest**

Random Forest significantly improved classification performance, reaching approximately 91% accuracy with an exceptional ROC–AUC of 0.993. The model demonstrated strong predictive capability across all five rating categories, as shown by its confusion matrix. Random Forest’s ability to model nonlinear interactions and high-order relationships allowed it to leverage both original and engineered features effectively. This robustness made Random Forest one of the top-performing algorithms in the project.

A graph with numbers and a number in squares

AI-generated content may be incorrect.A graph with a line graph

AI-generated content may be incorrect.

**9.4 XGBoost**

XGBoost achieved the highest overall performance, with an accuracy of nearly 92% and a macro-averaged ROC–AUC of ~0.99. Its gradient-boosting framework iteratively corrected mistakes from previous learners, producing highly refined decision boundaries. XGBoost consistently delivered excellent performance across all rating categories, including the more challenging extremes (A and F). Its strong predictive accuracy, stability, and ability to capture subtle feature interactions establish it as the best-performing model in this study.

A graph of a number of numbers and a number of blue squares

AI-generated content may be incorrect.A graph of a positive rate

AI-generated content may be incorrect.

**9.5 Summary of Findings**

Across all evaluated models, ensemble techniques—especially XGBoost and Random Forest—outperformed linear classifiers by a wide margin. These results highlight the presence of nonlinear and interaction-based patterns in Texas school performance data, which are better captured by tree-based methods. Baseline models offered transparency and interpretability, but the ensemble models provided the highest predictive power and most reliable classification of school ratings.

## 10. Conclusion

This project successfully developed a comprehensive machine-learning framework to predict Texas public school accountability ratings using academic, demographic, and distinction-based data. Through systematic preprocessing, thoughtful feature engineering, and careful model evaluation, we were able to uncover meaningful patterns that influence school performance across the state.

Our comparisons demonstrate that educational outcomes in Texas are shaped by complex, nonlinear interactions between achievement indicators, equity measures, and distinction recognitions. While baseline linear models such as Logistic Regression and LDA provided valuable interpretability, their predictive power was limited. In contrast, ensemble algorithms—particularly Random Forest and XGBoost—delivered substantially superior performance, with XGBoost emerging as the strongest model overall. Its high accuracy and outstanding ROC–AUC scores reflect its ability to capture subtle relationships among features and effectively distinguish between all five rating categories.

Beyond model performance, the project’s exploratory analysis offered important insights into geographic and demographic disparities. The interactive Power BI dashboards revealed strong regional clusters of A-rated schools, substantial variation between charter and non-charter campuses, and clear patterns in TEA distinction achievements. These findings underscore the value of combining statistical modeling with data visualization to support evidence-based educational decision-making.

Overall, this study demonstrates that machine-learning techniques—supported by engineered features and rigorous evaluation—can effectively predict school accountability ratings and highlight key drivers of academic success. The modeling framework and visual analytics developed here provide a strong foundation for future research, policy analysis, and statewide monitoring of school performance.

## 11. Future Work

Although this project produced strong predictive performance and valuable insights into Texas school ratings, several important extensions could further enhance the depth, accuracy, and real-world applicability of the analysis.

First, future research should incorporate tuition or funding-related information for each school, including per-pupil spending, local tax contributions, budget allocations, and additional financial resources available to campuses. Understanding the financial landscape behind each district and school could reveal new relationships between funding levels and accountability ratings. Integrating these financial indicators may help determine whether resource availability plays a measurable role in student achievement, equity outcomes, or distinctions earned.

Second, expanding the dataset’s TEA distinction categories with more granular details would improve the interpretability of model results. The current distinction indicators (e.g., ELA/Reading, Math, Science, Social Studies, Postsecondary Readiness) indicate whether a school earned recognition but do not capture the specific criteria, performance thresholds, or subject-level proficiency rates required to achieve them. Incorporating additional information—such as domain-specific sub-scores, student subgroup performance, or STEM-specific metrics—would allow for a deeper analysis of how distinctions contribute to school quality and how they correlate with statewide accountability ratings.

In addition to these enhancements, several other extensions could further improve the modeling framework:

* Include multiple years of accountability data to analyze trends over time and develop forecasting models.
* Conduct hyperparameter tuning for ensemble models to push performance beyond baseline configurations.
* Explore deep learning and advanced boosting methods for potential gains in accuracy and representation of learning.
* Perform subgroup fairness analysis to examine whether predictions differ across regions, school types, or demographic profiles.
* Integrate external datasets, such as teacher experience, attendance rates, graduation pathways, or community socioeconomic indicators, to enrich the predictive context.
* Expand Power BI dashboards to include real-time updates, drill-down funding comparisons, and interactive distinction profiles for districts and regions.

Together, these extensions would strengthen the predictive framework, improve transparency, and provide more comprehensive insights into the academic, financial, and structural factors shaping school performance across Texas.

## 12. References

Texas Education Agency. School Year 2022–2023 Statewide Accountability Ratings. Texas Open Data Portal, 2023,  
[*https://data.texas.gov/dataset/School-Year-2022-2023-Statewide-Accountability-Rat/nui6-x374*](https://data.texas.gov/dataset/School-Year-2022-2023-Statewide-Accountability-Rat/nui6-x374)*.*

Texas Education Agency. Distinction Designations Overview. 2023,  
[*https://tea.texas.gov/texas-schools/accountability/academic-accountability/performance-reporting/distinction-designations*](https://tea.texas.gov/texas-schools/accountability/academic-accountability/performance-reporting/distinction-designations)*.*

Trinh, Phuong, et al., Texas School Ratings Power BI Dashboard. 2025. [*https://app.powerbi.com/groups/me/reports/9e60b870-aeb8-4e5f-8300-9537faf1a9ae/b4f5805990120ceadc3b?ctid=5cdc5b43-d7be-4caa-8173-729e3b0a62d9&experience=power-bi&clientSideAuth=0&bookmarkGuid=3a8a485c-0811-4b34-860d-8e60ce788048*](https://app.powerbi.com/groups/me/reports/9e60b870-aeb8-4e5f-8300-9537faf1a9ae/b4f5805990120ceadc3b?ctid=5cdc5b43-d7be-4caa-8173-729e3b0a62d9&experience=power-bi&clientSideAuth=0&bookmarkGuid=3a8a485c-0811-4b34-860d-8e60ce788048)*.*