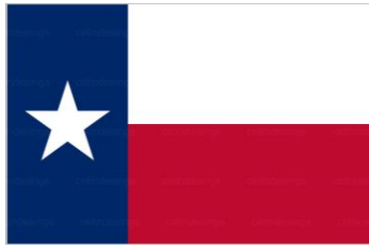
Avengers

District and School Ratings in Texas

Fall 2025

GROUP 2
Phuong Trinh
Harini Lingala
Sai Harshith Kondabathini
Mohamed Ahmed



Abstract

- **Goal 🎯:**
- Finding the best performing districts and schools in the state of Texas
- Aim is to help families with children looking to move to the Lone Star State
- **Approach:**
- ♟️ Compared the Overall Rating (A, B, C, D, F) of schools in the state given features such as counties, region, district, size of district, the distinctions of each school, and more
- 🛠️ Modeling: Logistic Regression, LDA, Random Forest, XGBoost
- 📊 Evaluation: Compare outcome of models (Accuracy, Precision, AUC)

📚 Why Districts and Schools Ratings?

👨‍👩‍👧‍👧 The education and growth of children is one of the most important things for parents

🏫 We analyzed the best schools given certain features to make it easy for parents to decide what school to enroll their kids in

Data Overview

- 🗂️ Collected from the Texas Education Agency (TEA) for the 2022-23 school year
- 📍 Consisted of over 10,000 schools and almost 1,000 districts

- 🔑 Key features:
  - District
  - County
  - Region
  - School type (Elementary, Middle, High)
  - Distinction Types
  - Charter
  - Overall Rating (target)

Workflow

Data Selection

↓

Data Preprocesing

↓

Feature Engineering

↓

EDA

↓

Train/Test Split

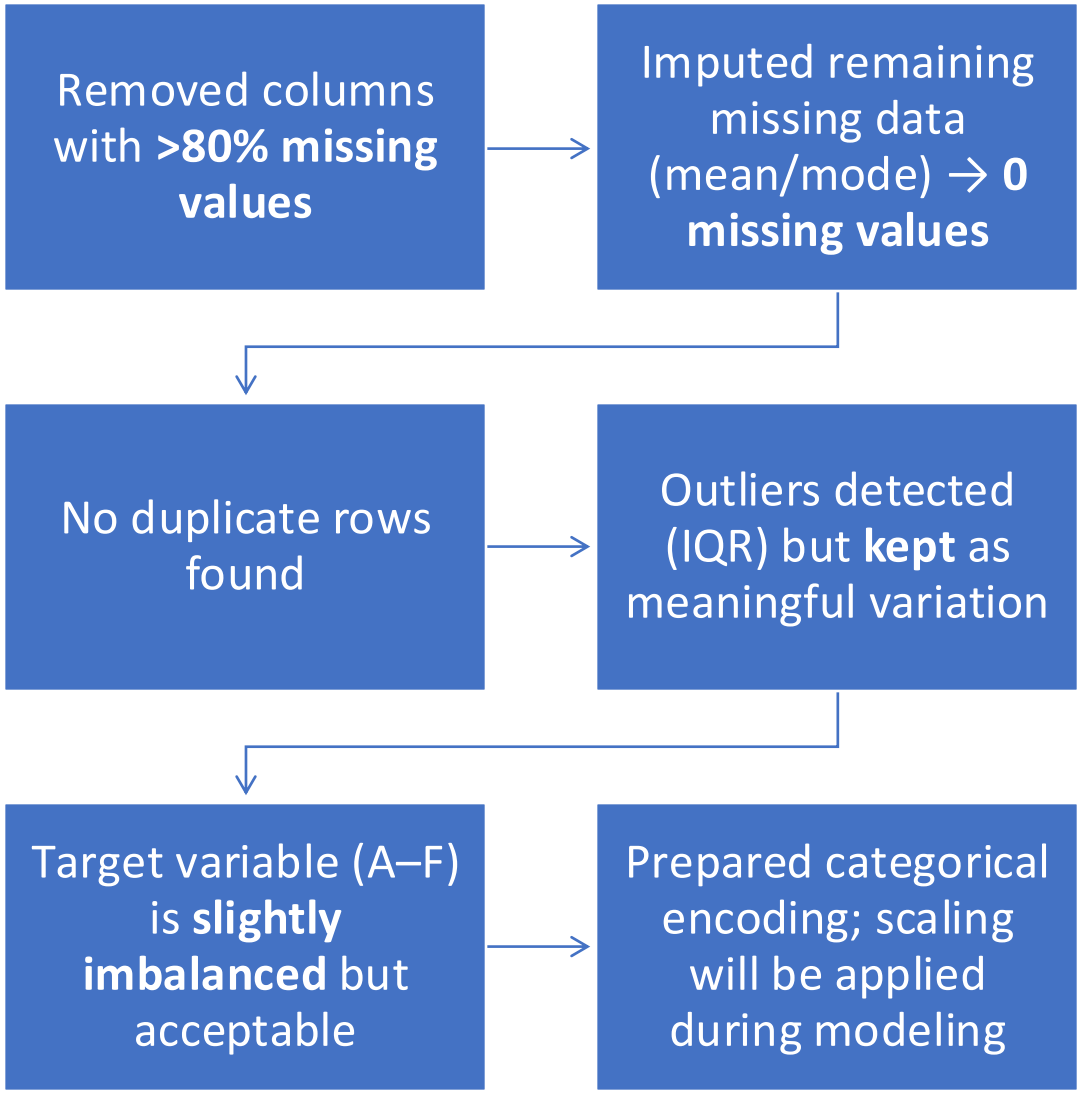
↓

Applying model
LDA/Logistic Regression
Random Forest/XGBoost

↓

Model evaluation

Data Preprocessing



Exploratory Data Analysis

Power BI

Feature Engineering

01

Performance Gap Index
• Measures inconsistency between Achievement/Progress vs Closing the Gaps
• Larger gap → lower school performance stability

02

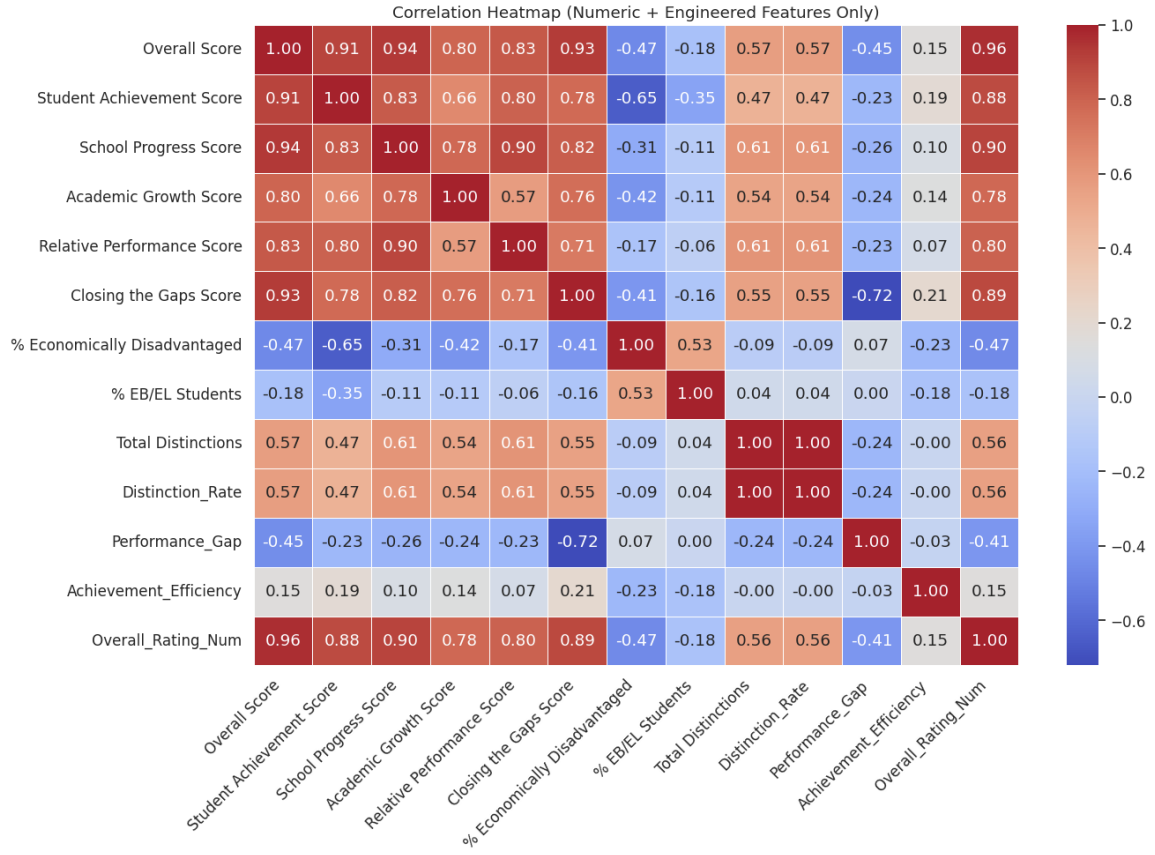
Achievement Efficiency Ratio
• Captures how well a school performs **relative to student size** and **economic disadvantage**
• Identifies schools "doing more with less"

03

Distinction Rate
• Converts total distinctions into a **normalized ratio** (0–1 scale)
• Higher rate → stronger academic recognition

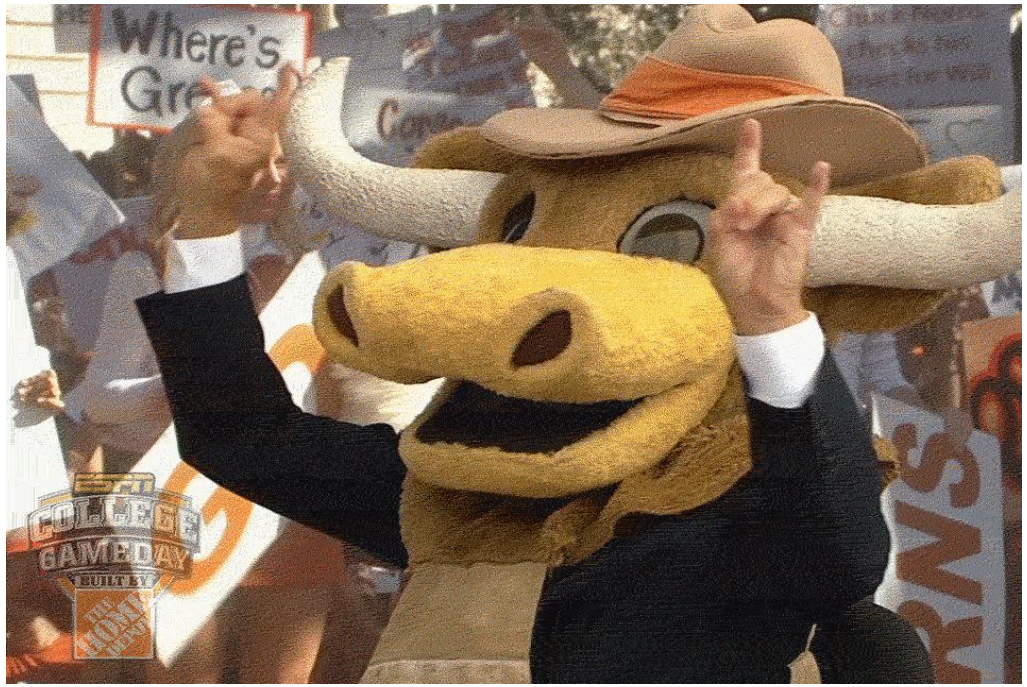
Feature Selection

- Chose essential **academic performance indicators** (School Progress, Academic Growth, Closing the Gaps)
- Added important **demographic + distinction features** (% Economically Disadvantaged, % EB/EL Students, Total Distinctions)
- Included **3 engineered features** (Performance Gap, Achievement Efficiency, Distinction Rate)
- Final features selected using the correlation heatmap:
  - **Strong correlation** with the target (Overall Rating)
  - **Low redundancy** across predictors
  - **High interpretability** for modeling + reporting



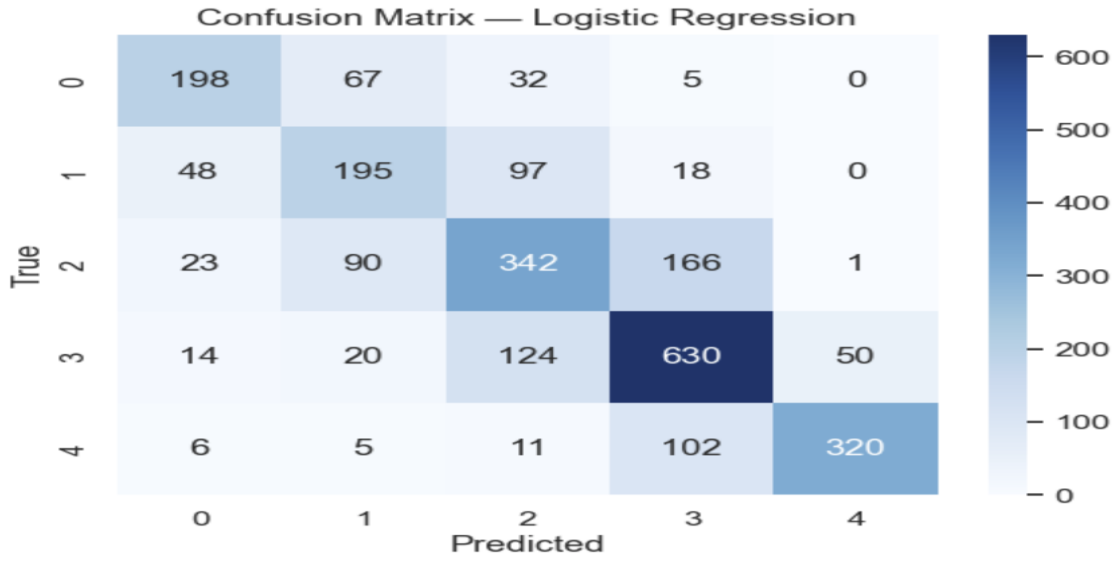
Machine Learning Models

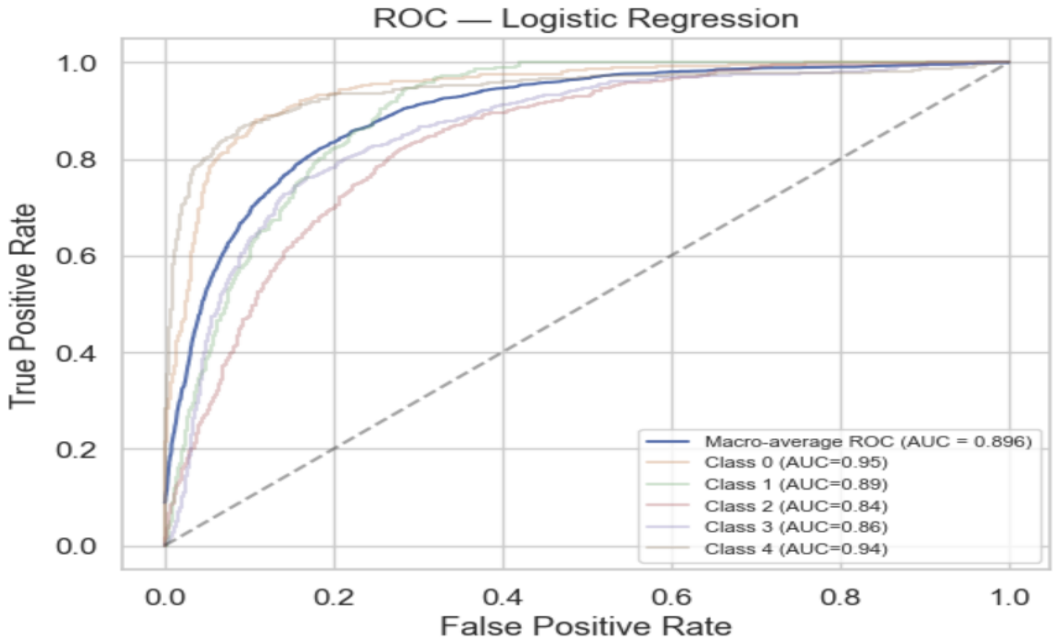
- Split data into training and testing sets (80/20)
- Scaled the data
- Confusion matrix
- Roc Curve
- Models Used:
  - LDA
  - Logistic Regression
  - Random Forest
  - XGBoost



Logistic Regression

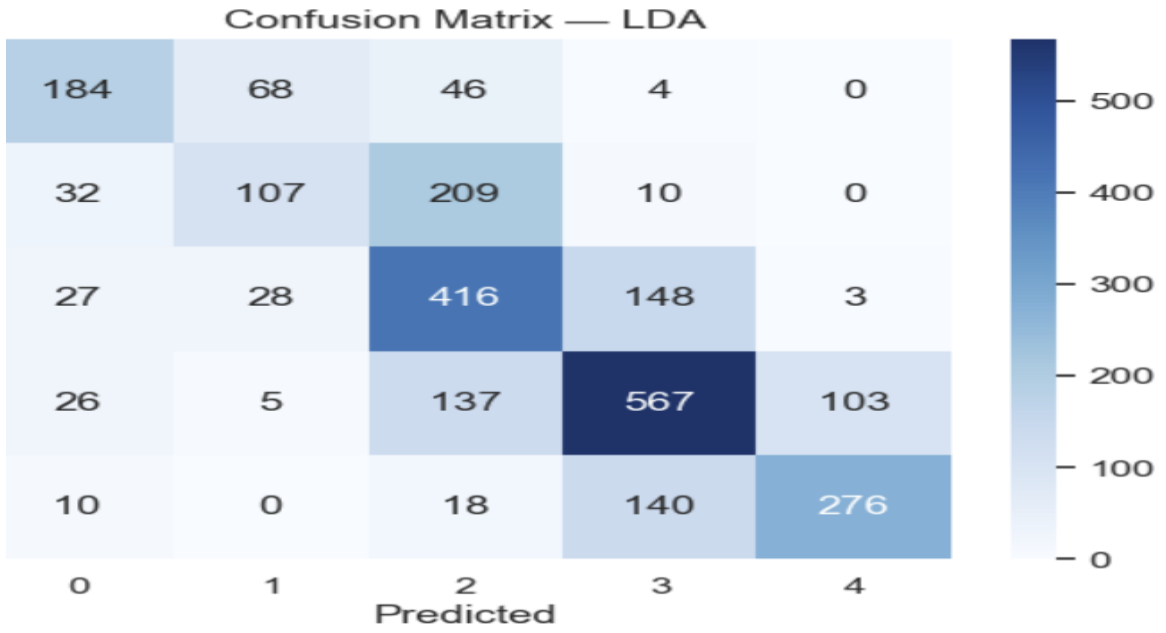
- Accuracy: 66%
- Works well for simple linear patterns
- Struggles to separate similar classes (1, 2, 3)
- AUC ~ 0.89(moderate class separation)

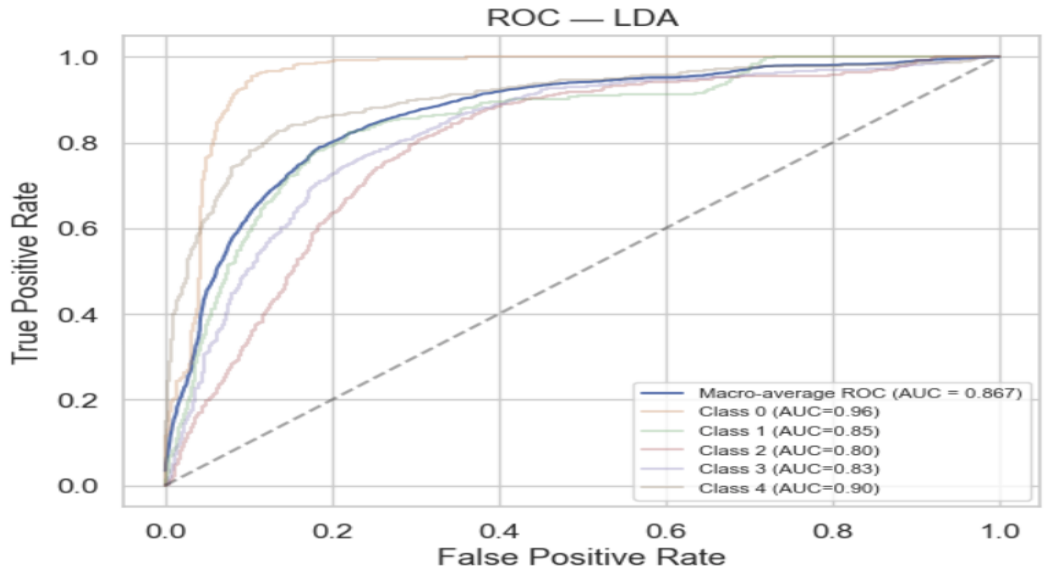




LDA

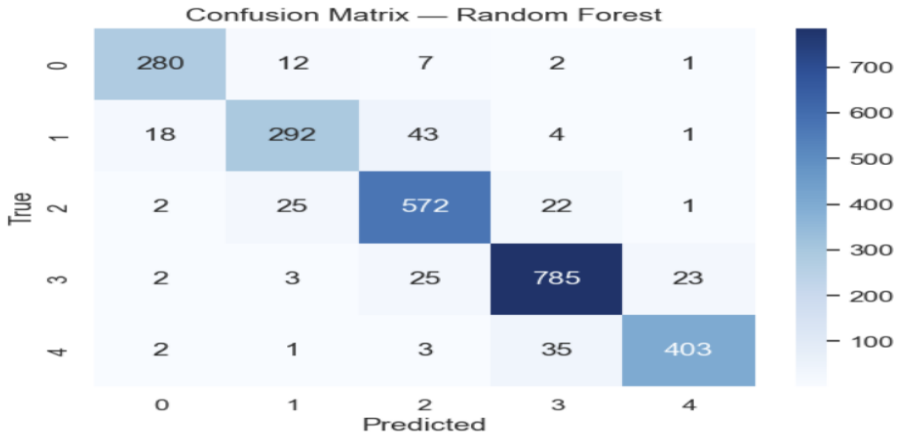
- Accuracy: 60%
- Struggles with non-linear boundaries
- Better at distinguishing extreme classes (0 & 4)
- AUC ~ 0.86

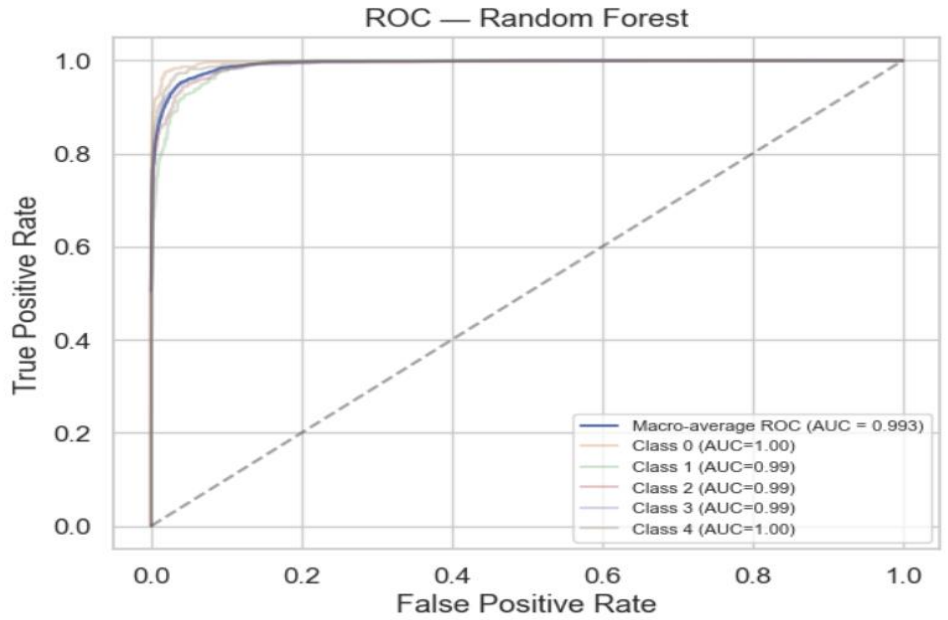




Random Forest

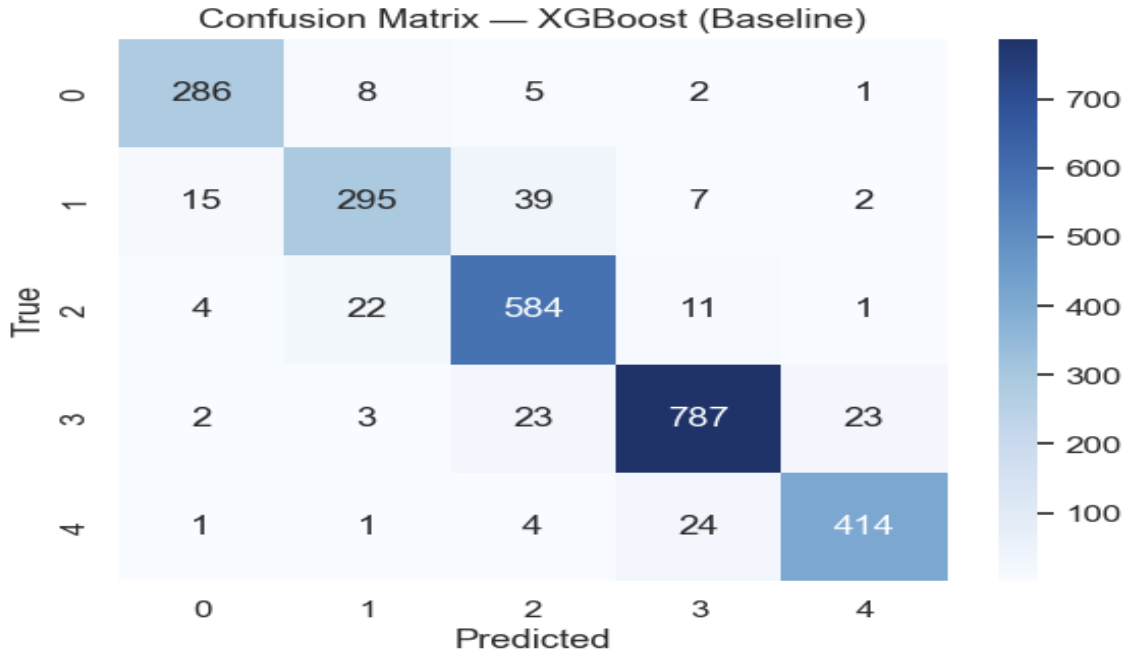
- Accuracy: 91%
- Very strong in capturing complex patterns
- Low misclassification across all classes
- AUC ~ 0.993

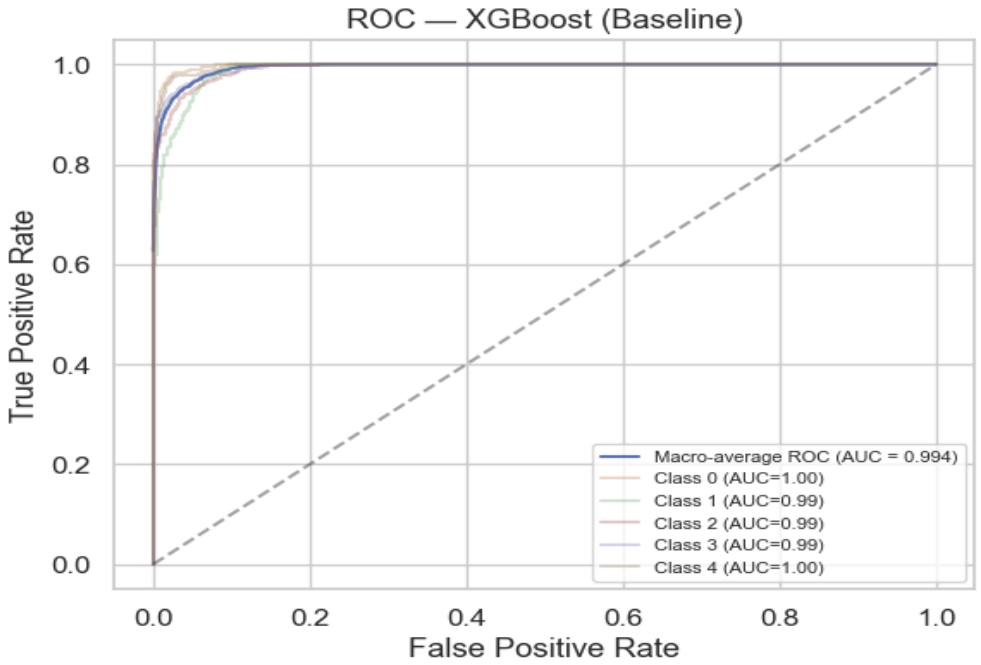




XG Boost

- Accuracy: 92%
- Most accurate model with very low errors
- Excellent for handling mixed-type features
- AUC ~ 0.994





WHICH MODELS IS THE BEST (CONCLUSION)

XGBoost performed the best overall

- Highest accuracy (~ 92%)
- Best macro F1-score (0.92)
- Highest ROC–AUC (~ 0.99)
- Consistently strong predictions across all classes
- Random Forest was the second-best performer
- Logistic Regression and LDA served as baseline models

FUTURE WORK

- Perform hyperparameter tuning for Random Forest and XGBoost to further improve accuracy.
- Use cross-validation to confirm the model's stability across different data samples.
- Explore additional algorithms such as SVM, KNN, or Neural Networks to compare performance.
- Conduct error analysis to understand which rating levels are most frequently misclassified.
- Try PCA or other dimensionality-reduction techniques to simplify the model if needed.
- Integrate the model outputs into the existing Power BI dashboard for automated model updates.
- We can try adding more data to improve the model performance.

Thank You