Results and Evaluations

Melody Goldanloo, Jenna Ramsey-Rutledge, Byron Selvage

Introduction:

This project aims to develop a model to predict student outcomes, such as graduation rates, from publicly available education data. In our last assignment, we built three exploratory models, a Random Forest Regressor, a Gaussian Mixture Model, and a Linear Regression Model. In this assignment, we aim to evaluate these exploratory models to determine the most suitable for use as our final model.

Random Forest Regression:

Our first exploratory model was a Random Forest Regression model. We used the metrics Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared Score (R^2) to evaluate our Random Forest Regression Model. The MAE was 0.1096. Since our output variable is only on a scale of 0 to 1, this is a relatively large average error. The RMSE was 0.1547. Since the RMSE is more sensitive to large errors, this suggests that the model may have inconsistent errors, where the model prediction deviates significantly from the true values for some observations. The R^2 score was 0.4574, which indicates that the model can only explain roughly 46% of the variance in the graduation rates. To further explore the model performance we plotted the true vs. predicted values of our test data. This graph is shown below.

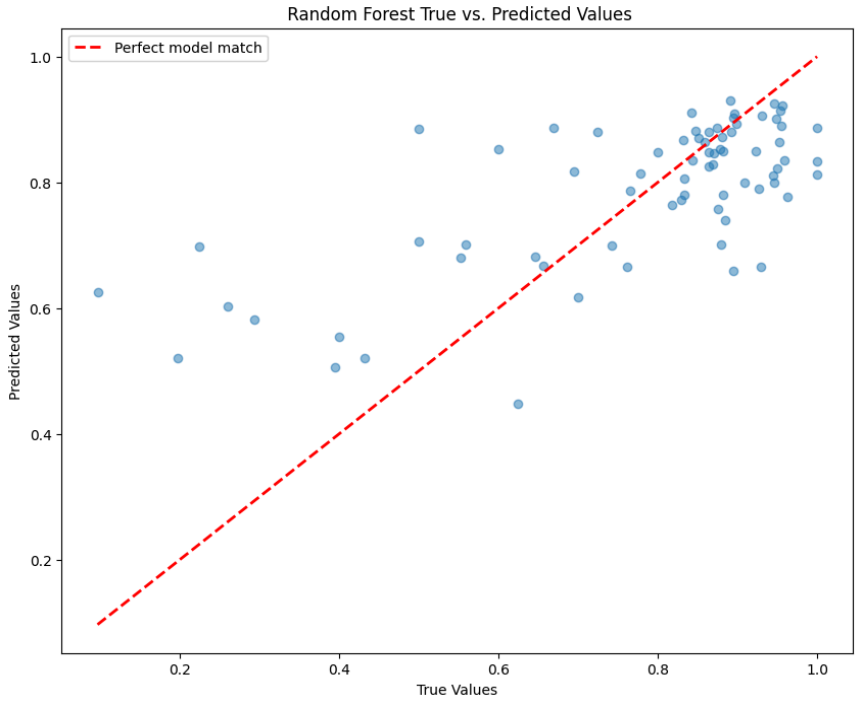

Figure 1: Random Forest True vs. Predicted Values

This plot shows that the Random Forest model struggles to accurately predict low graduation rates. This could suggest that the Random Forest model is too simple to capture the factors influencing low graduation rates. However, the true value of the Random Forest model is in its ability to evaluate feature importance. These feature importances are visualized in the figure below.

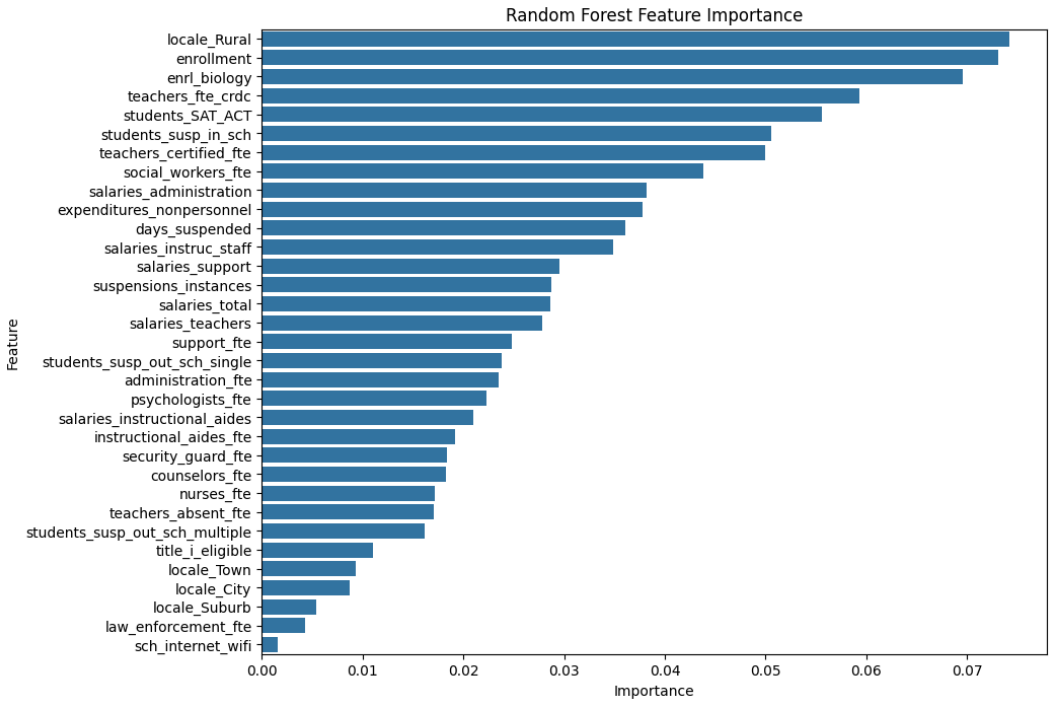

Figure 2: Random Forest Feature Importance

The pros and cons of the Random Forest model are summarized in the table below.

Pros	Cons
• Feature importance information • Resilient to overfitting • Captures non-linear relationships	• R-Squared Score: 0.4574 • Overestimation of true low graduation rates

Table 1: Random Forest Pros vs Cons

Gaussian Mixture Model:

For our second model, we used a Gaussian Mixture Model to cluster the data. We hoped that this would identify groupings within the data that would be helpful for predicting the graduation rate. We evaluated this model using the clustering metrics Log-Likelihood, Silhouette Score, Adjusted Rand Index, and Davies-Bouldin Index. The GMM model had a log-likelihood score of 20.2862. Since higher log-likelihood scores indicate better data fits, this score suggests that the model may fit the data reasonably well. The silhouette score was 0.1977. This indicates that the clusters may overlap or have

only limited separation. The adjusted rand score was 0.0265, which suggests that the clusters do not align with the known graduation rates data. The Davies-Bouldin index score was 1.8214. Although this score is somewhat relative, this higher value suggests that the model may have poorly separated clusters.

Since our data is high-dimensional, it was not helpful to visualize the performance of the GMM. The pros and cons of the Gaussian Mixture Model are summarized in the table below.

Pros	Cons
• Split the data into seven clusters (Changed scoring metric in GridSearch for slightly better performance) • Log-likelihood: 20.2862	• Silhouette Score: 0.1977 • Adjusted Rand Score: 0.0265 • Davies-Bouldin Index: 1.8214

Table 2: Gaussian Mixture Model Pros vs Cons

Linear Regression:

For our third model, we used Linear Regression to predict graduation rates. To try to optimize the model, we tested multiple techniques and variations. First, we used Principal Component Analysis (PCA) to reduce dimensionality, but the model's results: MAE = 0.136, MSE = 0.0329, R^2 = 0.254, Adjusted-R^2 = -0.3, indicated overfitting.

Next, we used Recursive Feature Elimination with Cross-Validation (RFECV) to select the 10 most impactful features. These features included enrollment, suspensions, and teacher FTE's (specifically, enrollment, students_susp_out_sch_single, suspensions_instnced, teachers_fte_crdc, teachers_absent_fte, counselors_fte, social_workers_fte, salaries_teachers, salaries_total, locale_Rural). These improved the model a little, with results: MAE = 0.146, MSE = 0.0347, R^2 = 0.214, Adjusted-R^2 = 0.0869.

With the same feature selection model, we experimented with a square root transformation and polynomial regression. The square root transformed model yielded: MAE = 0.157, MSE = 0.037, R^2 = 0.170, Adjusted-R^2 = 0.036. The polynomial regression model yielded: MAE = 0.141, MSE = 0.042, R^2 = 0.052, Adjusted-R^2 = -0.101. Overall, both led to reduced performance.

Below is a plot of the Actual vs. Predicted Graduation Rates using the Linear Regression Model with Polynomial Regression followed by its respective Residuals vs. Fitted plot. In the Actual vs Predicted plot, it is clear that the points do not follow a line as we would hope to see. The Residuals vs. Fitted plot allows us to see the variability in the data more clearly.

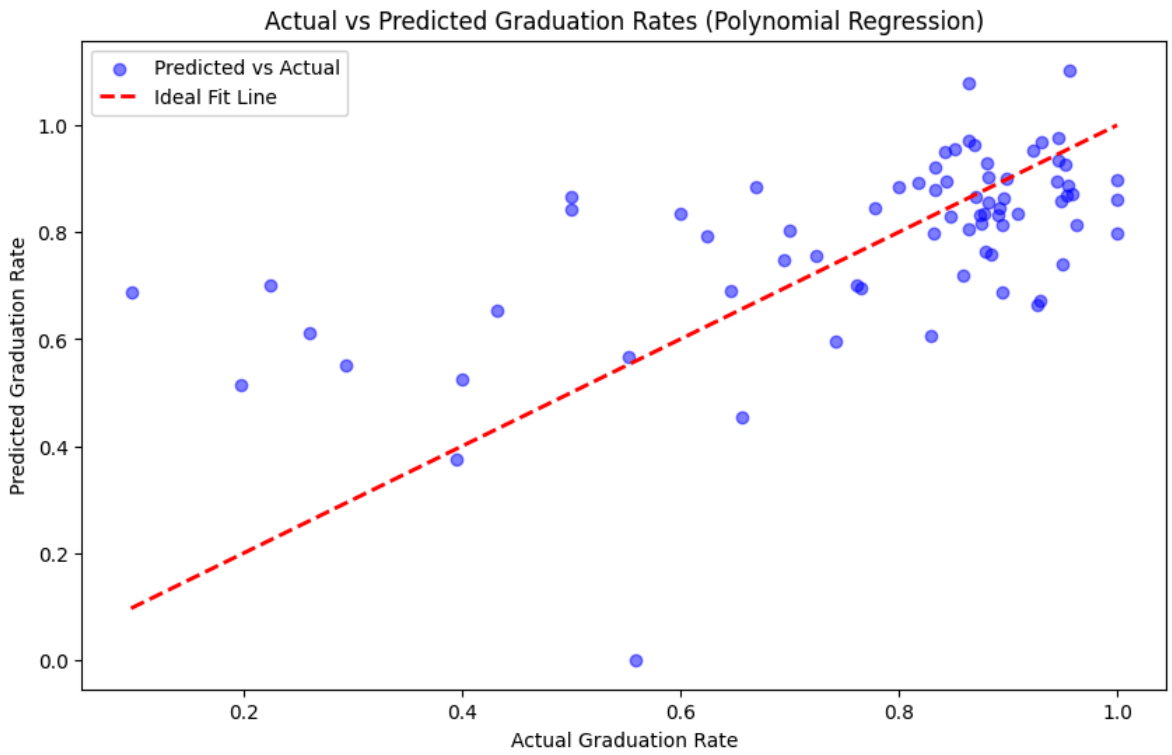

Figure 3: Polynomial Regression True vs. Predicted Values

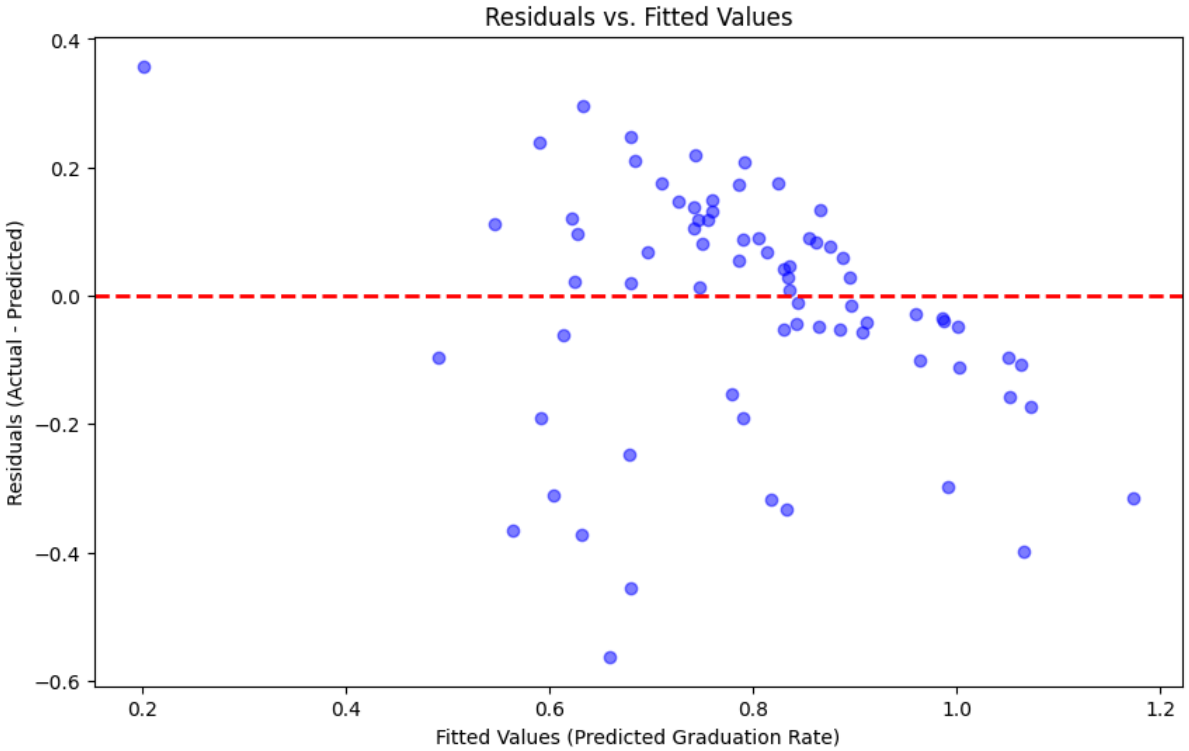

Figure 4: Polynomial Regression Residuals vs Fitted Values

Overall, these results suggest that linear regression is not well suited for this data, and a more complex model may better capture the relationships needed for accurate graduation rate predictions. The table below summarizes the pros and cons of the Linear Regression model.

Pros	Cons
• Simpler and more interpretable results • Computationally efficient • Easier feature selection	• Overall, poor performance on data (low R^2 values across all LRMs) • Sensitive to multicollinearity and outliers • Susceptible to overfitting (Negative Adjusted-R^2 values suggest overfitting)

Table 3: Linear Regression Model Pros vs Cons

Conclusion:

Based on the results of our exploratory models, the Random Forest Regressor model is the best option for predicting graduation rates from our data. The Random Forest model had an R^2 score of 0.4574 and an MAE of 0.1096. In contrast, the Linear Regression model was unable to produce an R^2 score higher than 0.254 or an MAE less than 0.136. The Gaussian Mixture model also performed poorly, unable to define clear clusters in the dataset.

Given these results, we will proceed with the Random Forest model and aim to further improve its performance. To do this, we aim to revisit the hyperparameter grid search process for finer tuning. We also aim to explore other versions of ensemble methods built on decision trees, such as Gradient Boosting.