Model Development

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. Our cleaned dataset focuses on high schools in Colorado and includes features such as enrollment, title I eligibility, teacher salaries, and total school expenditure. For this assignment, we chose to explore a Linear Regression model, a Random Forest Regressor model, and a Gaussian Mixture model.

Linear Regression:

We chose to use a Linear Regression model because it is flexible and has the ability to help determine feature importance. It is interpretable and easy to measure accuracy on. To build this model, we used GridSearchCV to determine whether ridge, lasso, elastic net, or regular linear regression model performed best. Each case had a mean square error of roughly 0.03. However, the Elastic Net model seems to have done slightly better than the rest. We intend to use the results from our other models to improve this model in the future.

Random Forest:

We chose to use a Random Forest Regression model because it evaluates feature importance in addition to providing a predictive output. Thus, even if the Random Forest model does not perform well on our data, we can use the information learned from it to improve our other models.

To build this model, we used GridSearchCV with 5-fold cross-validation to tune the hyperparameters 'n_estimators', 'criterion', 'max_depth', 'min_samples_split', 'min_samples_leaf', and 'max_features'. This created a model with a mean absolute error of 0.1096, mean squared error of 0.0239, root mean squared error of 0.1547, and R-squared of 0.4574. Since our target variable is on a scale of 0-1, this model performs reasonably well. However, another model may better fit the relationships of our data.

Regardless, the true value of this model comes from the feature importance, which tells us that features like the number of students enrolled in biology classes are important while features like Wi-Fi availability are unimportant. We will use these results to improve our other models in the future.

Gaussian Mixture Model:

We chose to use a GMM to test if there were clusters in the data. The goal was to use the model to decipher if there were different levels of graduation rates dependent on the features we collected. The use of this model was primarily exploratory in nature.

To build the model, we used GridSearchCV once again to determine the best hyperparameters. The output was “Best Parameters: {'covariance_type': 'full', 'n_components': 2}.” So, the model was created with 2 components and a covariance type of full. The results were difficult to interpret, which means this likely was a doomed endeavor. The silhouette score was 0.1115 and the fact that the most accurate number of components is 2 indicates that the data is not easily sorted into clusters based on graduation rates.