Final Project - Segment 4

By Pa Lor, Eric Sanders, and Sue Yang

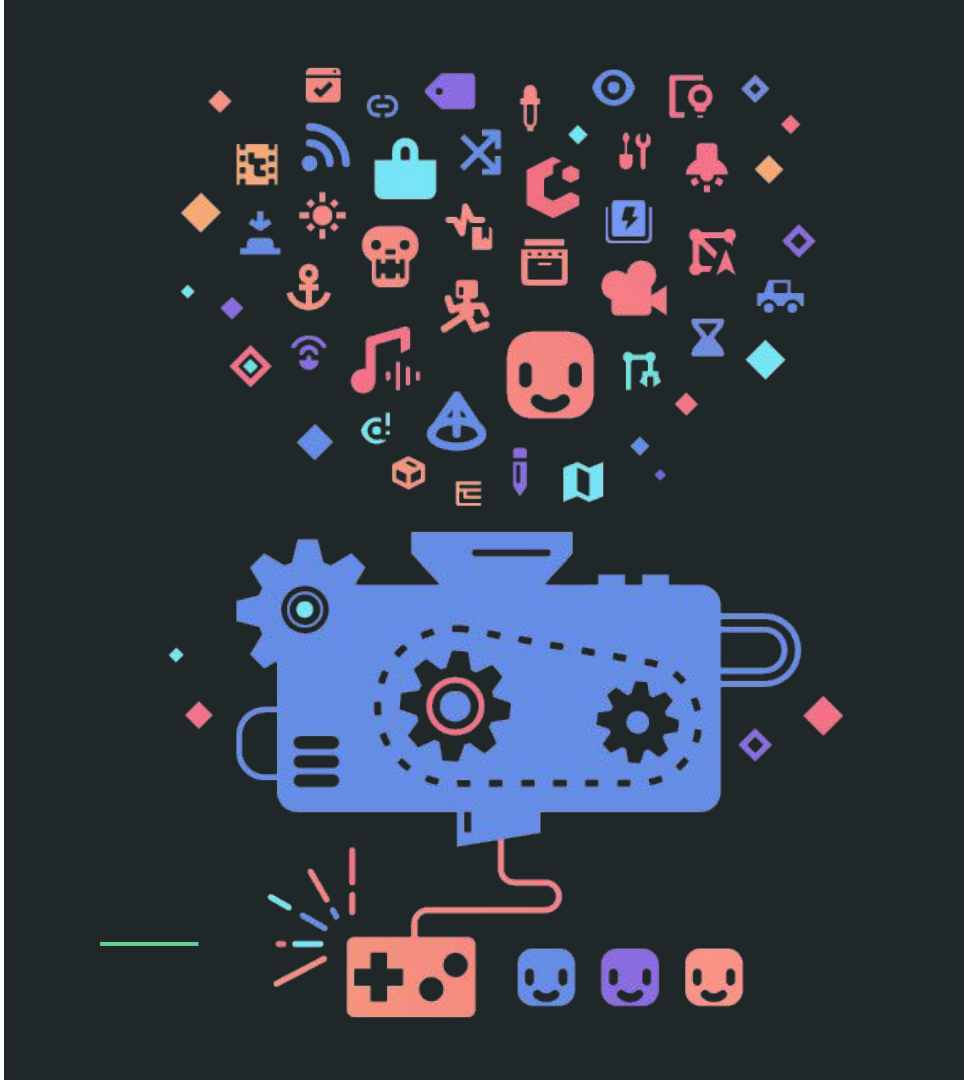
https://github.com/yangsue47/School_Is_Cool

Impact of Environmental Factors on Student Test Scores

As the pandemic resides and schools return to some level of normalcy, academic performance data, specifically data on environmental factors contributing to a student's academic test scores, will be crucial for evaluating how to get students back on track academically.

The dataset used for this project was pulled from Kaggle and linked here.

Using a machine learning model, we will identify which features have the largest impact on predicting a student's growth in test scores.



Questions we hope to answer:

Based on environmental factors such as school setting, school type, teaching method, classroom size, and socio-economic status(indicated by lunch status), can we predict whether a student will have high or low growth in their test scores (pre-test compared to post-test scores)?

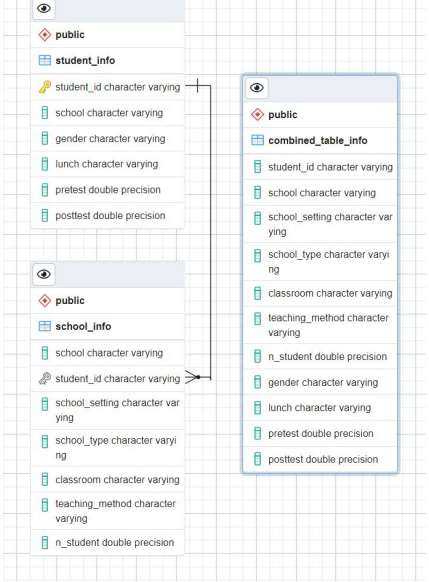
Which of the factors assess are the most important in determining a student's growth?

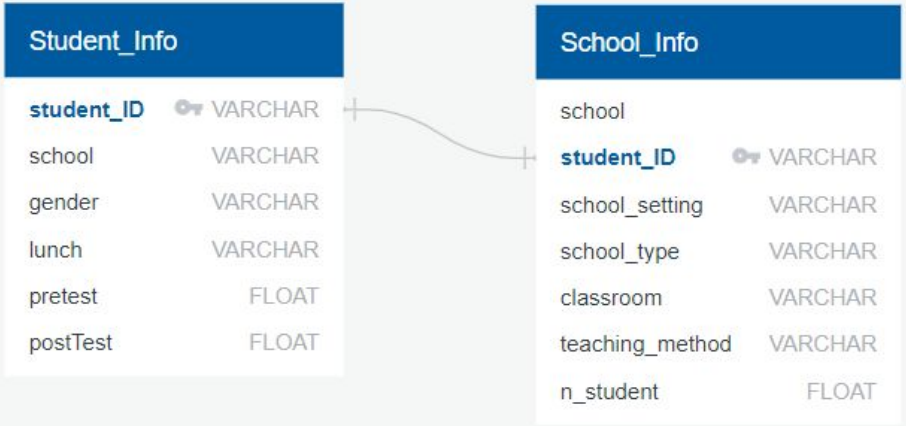
Data Exploration & Analysis Phase

- Very little cleaning was needed for our dataset.
- Descriptive statistics were generated to understand makeup of dataset. This was completed for the whole dataset and for each school setting (urban, suburban, and rural).

Data Exploration & Analysis Phase - cont'd

- To measure growth in test scores, we calculated % change in score from pre- to post-test. Then used that data to determine if a student experienced high or low growth.
- Based on the mean for % change in test scores (about 18.63), we decided to go with 18.63 points as our cut-off score for determining our high and low growth categories.





Technologies & Tools

- Python
- VS Code/Jupyter Notebook
- Tableau
- Postgres SQL
  1) Student information
  2) School information

Machine Learning Model

Model Choice: Random Forest

Some limitations of our model choice includes:

- Small data set
- Training time and resource allocation (in the event of a large dataset)

Some benefits include:

- Feature importance ranking
- Works well with categorical & continuous variables
- Robust to outliers

Machine Learning Model - Continued

- Minimal preliminary data preprocessing needed
  - Creation of Low and High Growth groups for Y variable
  - Convert categorical features into numerical encoders
- X = school setting, school type, teaching method, number of students, gender, and lunch
- Y = Low or High Growth
- Balanced split between Low Growth and High Growth groups

Model - Confusion Matrix & Accuracy Score

```
from sklearn.metrics import balanced_accuracy_score
# Calculated the balanced accuracy score
balanced_accuracy_score(y_test, predictions)
```

```
0.7383165049638989
```

	Predicted High_Growth	Predicted Low_Growth
Actual High_Growth	171	85
Actual Low_Growth	53	224

Classification Report

```
# Print the imbalanced classification report
from imblearn.metrics import classification_report_imbalanced

print(classification_report_imbalanced(y_test, predictions))
```

```
                   pre       rec       spe        f1       geo       iba
sup

High Growth       0.76      0.67      0.81      0.71      0.73      0.53
256
 Low Growth       0.72      0.81      0.67      0.76      0.73      0.55
277

avg / total       0.74      0.74      0.74      0.74      0.73      0.54
533
```

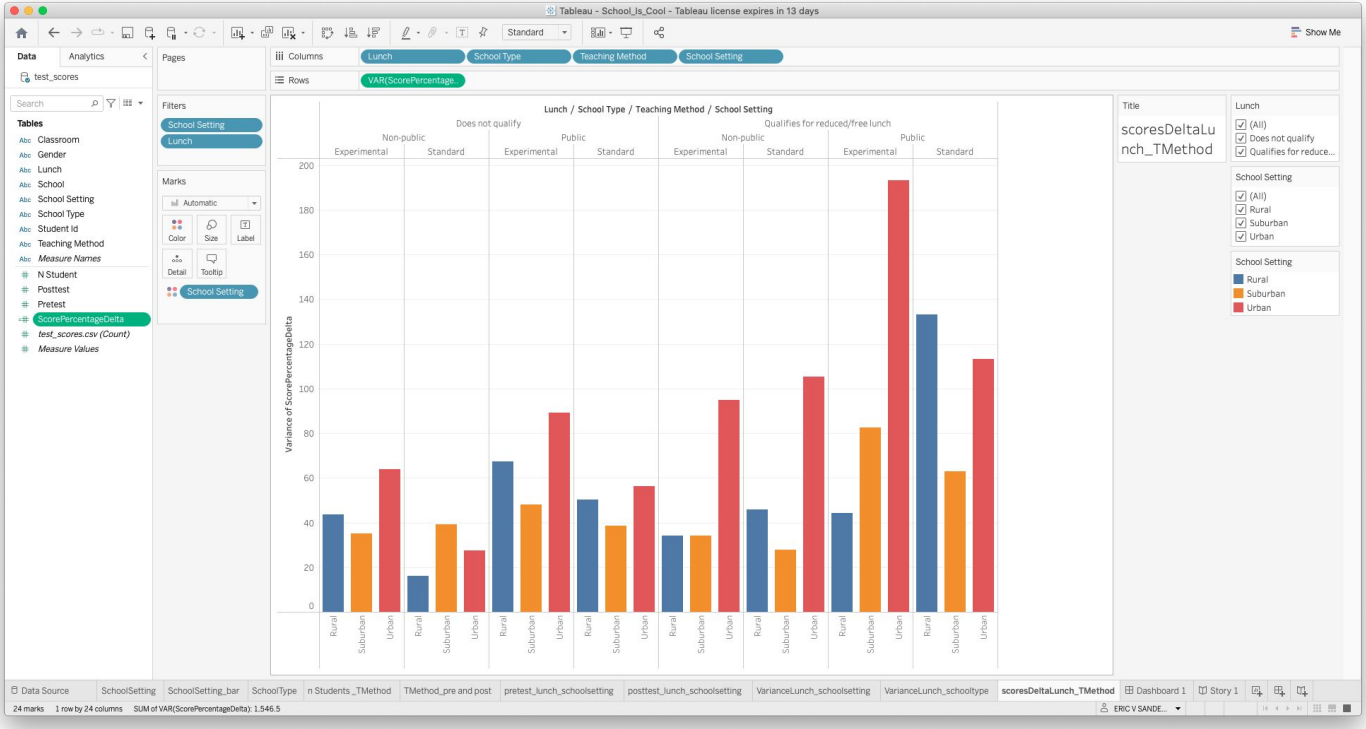
Feature Importance

```
# List the features sorted in descending order by feature importance
importances = rf_model.feature_importances_
sorted(zip(rf_model.feature_importances_, X.columns), reverse=True)
```

```
[(0.31476418876454954, 'n_student'),
 (0.16731796831573242, 'teaching_method_Standard'),
 (0.1470515559602089, 'teaching_method_Experimental'),
 (0.09002074433037352, 'lunch_Does not qualify'),
 (0.08985840288336173, 'lunch_Qualifies for reduced/free lunch'),
 (0.04475929103266995, 'school_setting_Suburban'),
 (0.034399440534362204, 'school_setting_Rural'),
 (0.02594011148049291, 'school_setting_Urban'),
 (0.021835190760545017, 'school_type_Non-public'),
 (0.021775567092433255, 'gender_Female'),
 (0.02154581650641328, 'gender_Male'),
 (0.020731722338857l7, 'school_type_Public')]
```

The number of students in the classroom impacts a student's growth factor the most.

Exploratory Visualizations

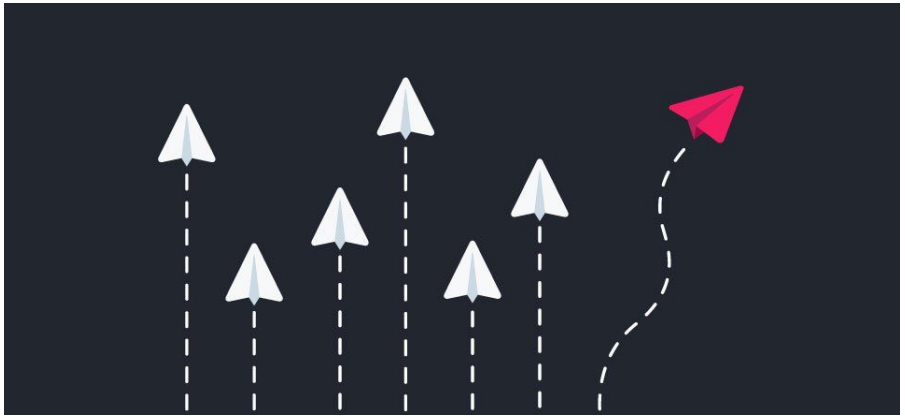


Final Dashboard

Recommendations For Future Analysis

- More robust dataset
- Correlations between pre- and post-test scores & predictions
- Further analysis on features and why they may/may not be important in test score growth
- Feature importance based on specific populations

Project Learnings - What Would We Do Differently?



- Establishing a better framework to guide our selection of a dataset
- Explore additional machine learning models during initial phases of project
- Use multiple branches and merge into the main branch

Any questions?

Thank you!