Final Project Progress Report 2

Group name: Ascending

Group members: Jeanette, Kunjian, Shirley, Carrie

GitHub: https://github.com/macs30122-winter25/final-project-ascending.git

Project Description

Originally focused solely on Illinois, our project now examines school performance in commuting zones across Wisconsin, Illinois, and Indiana. By broadening the geographic scope and sample size, we aim to enhance the generalizability of our findings.

At this stage, we included a sentiment analysis of user reviews on GreatSchools.org to capture students' and families' personal experiences and attitudes toward the school. Specifically, using a VADER model, we adopted the sentiment compound valence matrix for each school's comments. It has four components: a positive score, a negative score, a neutral score, and a compound score, a combination of the previous three scores. We finished the sentiment analysis and included some descriptive data and visualizations. See src/collect_reviews.py and notebooks/reviews_analysis.ipynb.

Data Sources

Firstly, we have made significant progress in data collection. We aim to examine school performance in commuting zones across Wisconsin, Illinois, and Indiana. We have scraped all the data from GreatSchools.org at the school level, including all the sections on school performance, demographics and geographic information on the webpage. See src/collect_schools.py and ratings_collector.ipynb for the codes, and data/greatschools/schools_data_all_states.csv for the scraped geographical and identification data. data/greatschools/scraped_ratings_all_states.csv contains quantitative data that are shown on GreatSchools, such as student demographics, and academic

performance scores. Secondly, we also scraped review pages for all the schools, see data/greatschools/reviews_data_all_states.csv.

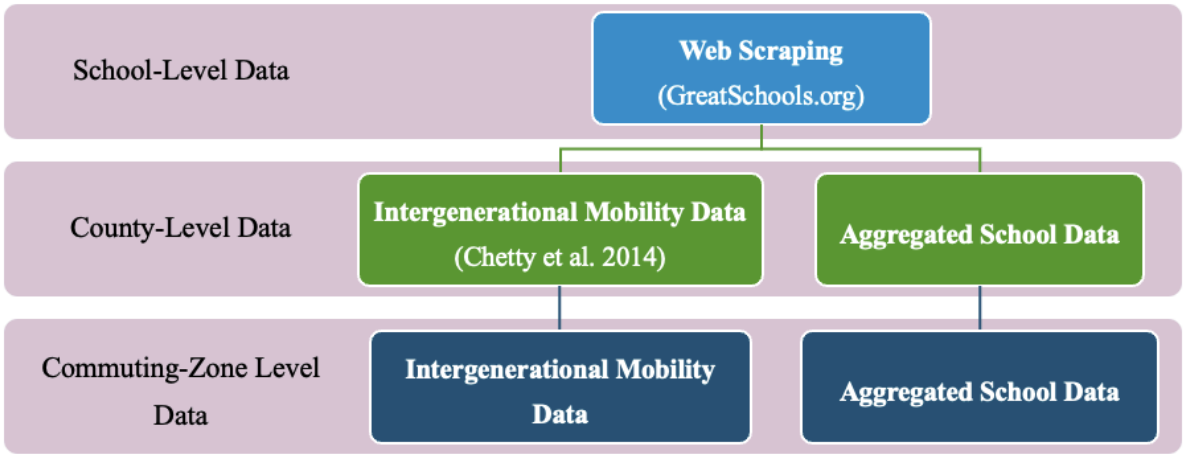
We obtained data sets with mobility rates at the county level and commuting level in the U.S. from a previous research, *Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States* by Chetty et al. (2014). See data/intergenerational_mobility.xls.

Before running the machine learning models and regression, we will need to collect governmental data which includes the Tract-FIPS and economic and demographic variables. We hope to control for the economic and demographic variables to better demonstrate the relationship between school performance and intergenerational mobility.

Data Cleaning/Wrangling

We finished the construction of the converter, which is designed to scale our independent variable, educational performance at the school level, to a commuting zone level, which contains the intergenerational mobility data. Each school's longitude and latitude are also scrapped from GreatSchool. We matched the schools' geographical information with this area's census tract ID (Tract_FIPS). Since this ID is an official label, we can aggregate the school performance with the commuting zone. On a side note, we also convert county data, which is used to perform robustness analysis to examine results on different levels of the data to check for consistency. See src/converter.py for the codes, and data/geo for data.

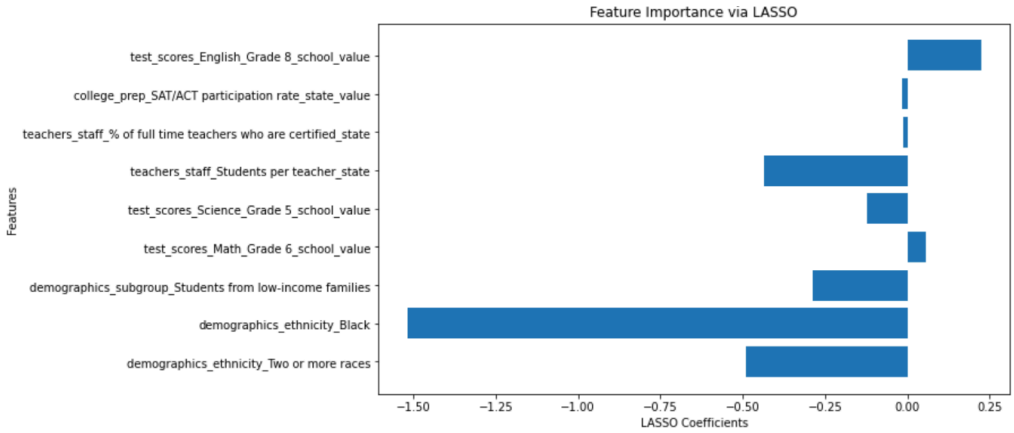
We also developed an aggregation strategy featuring our scraped data from GreatSchool and census data. We merged datasets containing school-level data with the school-county-commuting zone results from converter.py. In this way, we were able to create two csv files. Each file contains school features that were scraped from GreatSchools and mobility rate at that level.



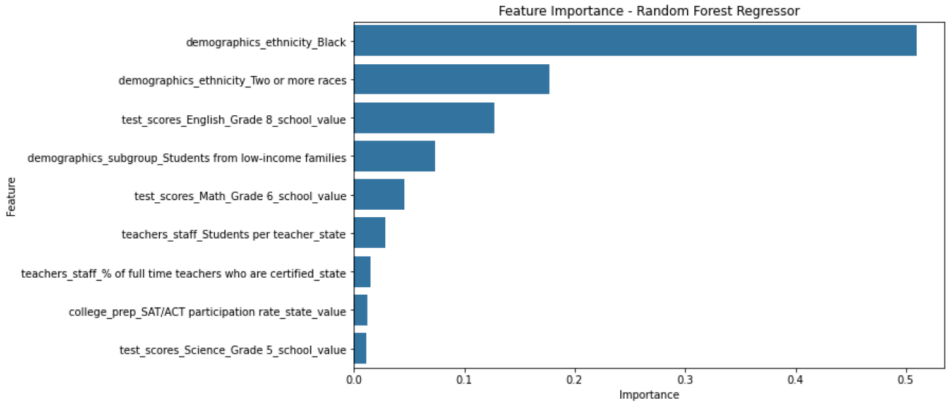
Data Analysis

We have refined our supervised machine-learning models. The codes for OLS and random forest models are ready, as is a function for residual analysis. In addition, we have several plots for model visualization. See src/regression.py for more details. We performed data analysis on both the county level and the commuting zone level. We used machine learning models to identify the most influential factors affecting mobility.

We first ran LASSO to reduce dimensionality and set irrelevant features to zero.



Then, we trained a Random Forest model with the rest of the features and ran OLS. This allowed us to see feature importance.



OLS Regression Results

Dep. Variable:	Absolute_Upward_Mobility	R-squared (uncentered):	0.999
Model:	OLS	Adj. R-squared (uncentered):	0.998
Method:	Least Squares	F-statistic:	879.7
Date:	Wed, 26 Feb 2025	Prob (F-statistic):	7.36e-17
Time:	23:06:37	Log-Likelihood:	-47.216
No. Observations:	25	AIC:	118.4
Df Residuals:	13	BIC:	133.1
Df Model:	12		

	coef	std err	t	P>|t|	[0.025	0.975]
demographics_subgroup_Students from low-income families	-6.682e-09	2.41e-09	-2.769	0.016	-1.19e-08	-1.47e-09
County_FIPS	0.0024	9.02e-05	26.151	0.000	0.002	0.003
test_scores_English_Grade 3_school_value	0.0710	0.073	0.974	0.348	-0.087	0.229
test_scores_English_Grade 4_school_value	0.0848	0.043	1.976	0.070	-0.008	0.178
test_scores_English_Grade 5_school_value	-0.0926	0.048	-1.939	0.075	-0.196	0.011
test_scores_Math_Grade 3_school_value	0.0201	0.070	0.288	0.778	-0.131	0.171
test_scores_Math_Grade 5_school_value	-0.0946	0.073	-1.298	0.217	-0.252	0.063
test_scores_Math_Grade 6_school_value	0.1449	0.071	2.040	0.062	-0.009	0.298
test_scores_Science_Grade 5_school_value	0.0479	0.162	0.296	0.772	-0.302	0.398
test_scores_Social Studies_Grade 11_school_value	-0.0373	0.116	-0.323	0.752	-0.287	0.212
test_scores_Math_Grade 7_school_value	0.0672	0.067	0.997	0.337	-0.078	0.213

We analyzed our scraped data to find the relationship between educational performance and intergenerational mobility data. We need to do this: More detailed and thorough analysis is needed. We will visualize the states with 3D aerial plots.