Analyzing socio-economic factors that contribute to high school student dropout rates in California

Arcy Flores, Haoxin Luo, Xiangnan Zhou, Xuehui Yang, Yiqiu Shen

Highlights

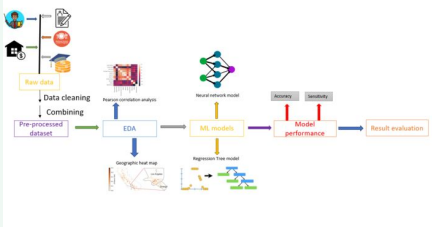
- High school dropout rates are positively influenced by free lunch eligibility and under-qualified staff, but negatively influenced by qualified staff
- Mitigation measures are needed to help economically-disadvantaged students and schools with limited teacher resources
- Dropout rates in high school is a complicated multi-layer socio-economic problem that requires further investigation

Background

California has the largest student population in the United States and contains one of the most diverse populations in the nation, making it essential to understand the challenges and barriers faced by students in achieving high school graduation.

Model

Decision Tree Models: Important features are shown near the root. Beta Regression Models: Reveals predictor-dropout rate relationships. Neural Network Models: Uncover feature importance through weight parameters and visualization techniques.



Data

We collect relevant data from several sources including the California Department of Education, California State Geoportal, U.S. Census Bureau, and ACS-ED. We focus on relevant social and economic features, the important ones are shown below. The FTE represents full-time-equivalent staff.

Category	Features
Social	**Enrollment, number of clear/other FTE**
Economic	Funding, median household income, **free lunch eligibility**

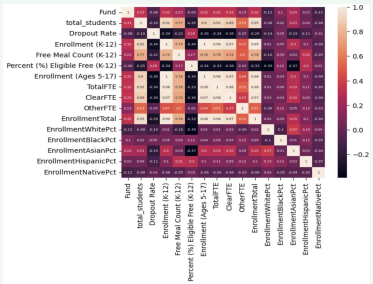
Note: school level (bold), district level (normal)

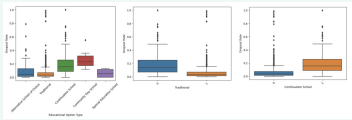
Spatial Visualization of High School Dropout Rates in California.

Data cleaning

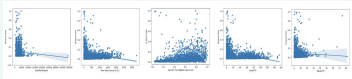
- Process individual datasets and store data in dataframe formats with unique rows indexed by school ids.
- Filter schools that have grades 9-12
- Join features together and drop rows with missing values
- Final clean data size: 1498

Findings

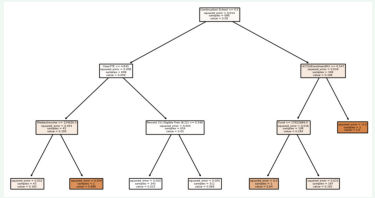


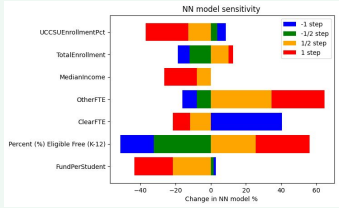


Seaborn boxplots show significant differences in dropout rates across levels of education option type. We focus on continuation and education school levels as they are correlated with each other



Heat maps displaying correlation score of all features, computed using Pearson's coefficient. Variables that exhibit a moderate relationship with dropout rates include total student count, alternative school of choice, continuation school enrollment, and traditional school attendance. This is further provide by the regression plots





	coef	std err	z	P>|z|	[0.025	0.975]
FundPerStudent	-3.925e-06	1.13e-05	-0.346	0.729	-2.62e-05	1.83e-05
Percent (%) Eligible Free (K-12)	-0.7086	0.567	-1.250	0.211	-1.820	0.402
ClearFTE	-0.0237	0.013	-1.880	0.060	-0.048	0.001
OtherFTE	0.0086	0.021	0.407	0.684	-0.033	0.050
MedianIncome	-6.335e-06	5.53e-06	-1.145	0.252	-1.72e-05	4.51e-06
totalEnrollment	8.516e-05	8.06e-05	1.056	0.291	-7.29e-05	0.000
UCCSUEnrollmentPct	-1.2919	0.375	-3.447	0.001	-2.027	0.557
NCLBEnrollmentPct	0.0049	1.167	0.004	0.997	-2.282	2.292

According to the trained models, the three most significant features are continuation school enrollment, clearFTE, and UCCSU enrollments. Other important features include median income, percent eligible for free meals (K12), and fund.