Learn More, Spend Less

A linear regression analysis of college tuition prices in the U.S.

Max Harrington
Metis Linear Regression Module

Introduction

Motivation:

- College prices have risen 1,416% in the past 40 years, outpacing inflation.
- U.S. families must decide if a formal education is worth the significant debt

Objective:

- Create a tuition estimator which allows families to see if a school is over/undervalued.

Project Goal:

- Estimate an appropriate tuition price based on a school's various metrics, and identify the relative importance of each feature.

Data was scraped from National Center for Education Statistics, a branch of the U.S. Department of Education.

Metrics Used

- **Tuition**
- **Enrollment**
- **Student/Faculty Ratio**
- **Room and Board**
- Books/Other
- **Financial Aid/Loans**
- **Enrollment Demographics**
- Applications Demographics
- **Test Scores**
- **Retention and Graduation Metrics**
- Degree Types
- **Athlete Count**
- **Crime Rates**

After scraping, 47 initial features were found (many categorical)

Gold features were ultimately used

The data initially included all U.S. Colleges offering bachelor's degrees, 2,357 schools.

TOOLS

Beautifulsoup

Scraped data from
nces.ed.gov/collegenavigator/

Pandas/Numpy/ Regex

Used to clean, prepare, and explore data

SciKit-Learn and statsmodels

Used to build and interpret linear models

Matplotlib and Seaborn

Create visualizations of results

01

Data Preparation

- Scraped and cleaned relevant data
- Removed/formatted data for modeling
- Created dummy variables for categorical features

02

Modeling

- Used Linear, LASSO, Ridge, and Polynomial Regression for analysis
- Used cross validation and scaling to improve accuracy and explainability, respectively.

03

Feature Engineering

- Found interactions between population based features
- Modeled polynomial relationships
- Eliminated or combined collinear variables

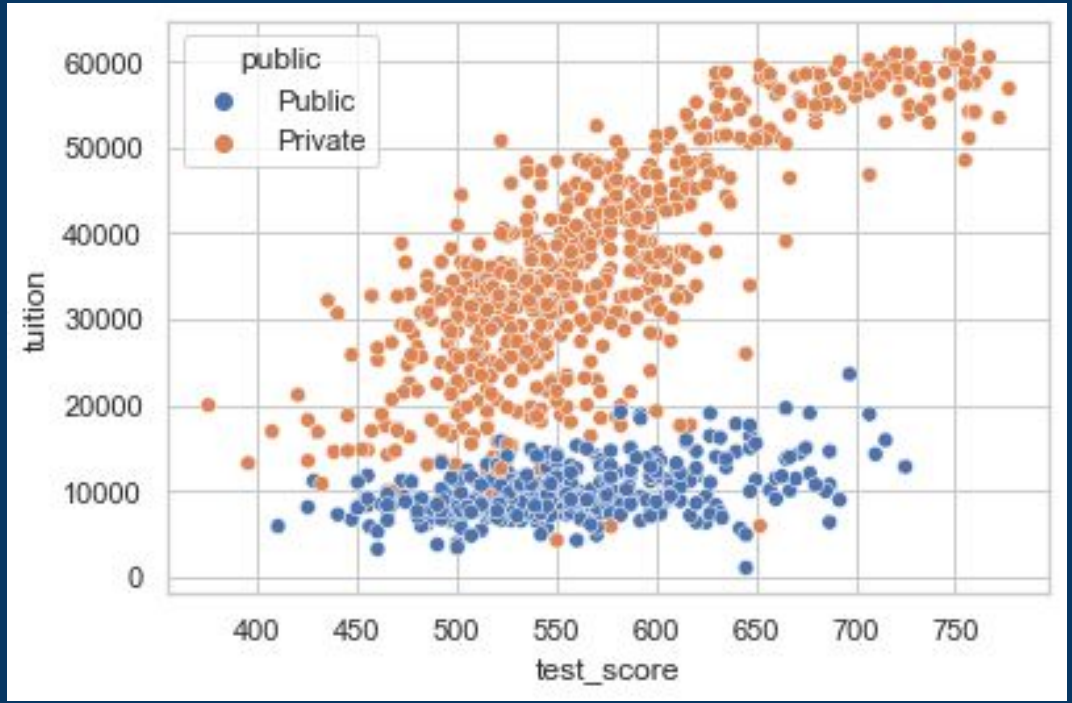
04

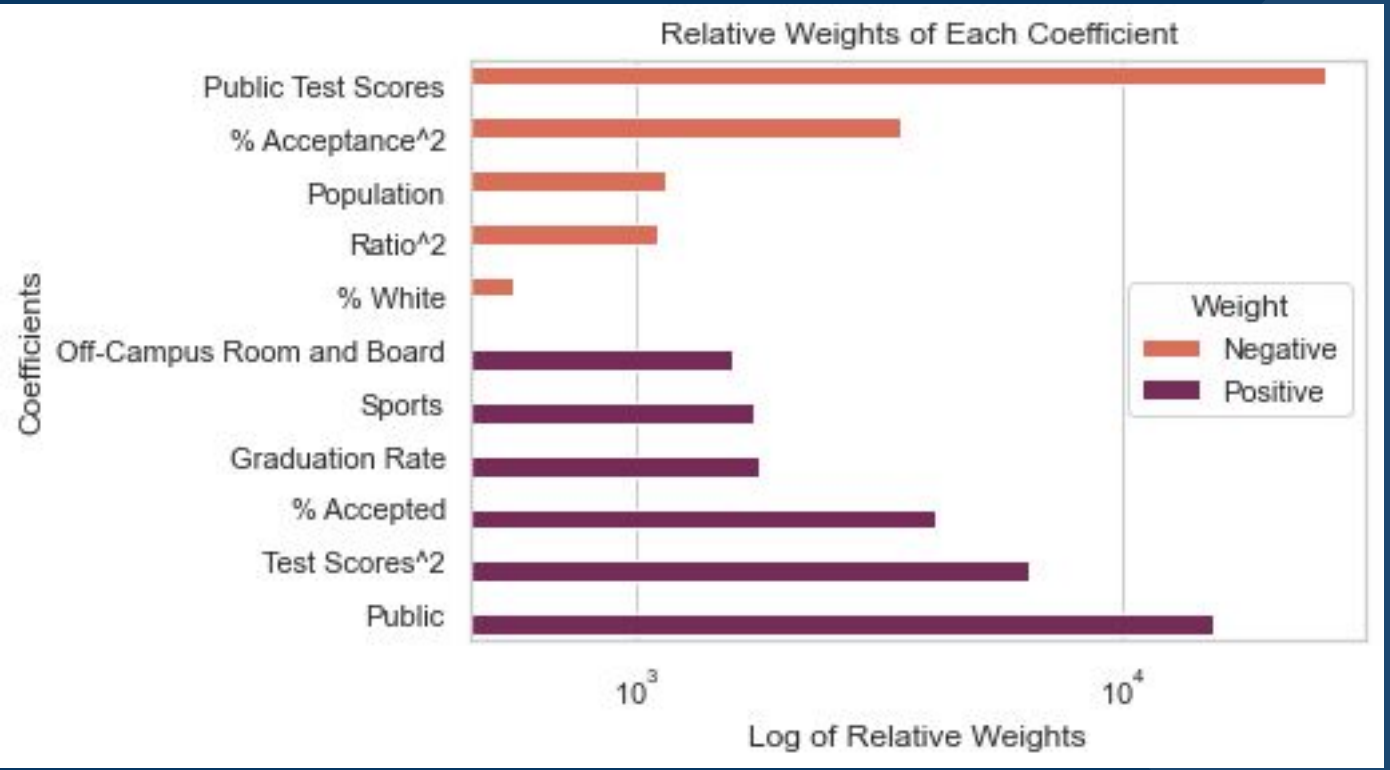
Graph + Analysis

- Created multiple visualizations of key relationships
- Identified strongest relationships, coefficients and models.

Public vs. Private Test Score

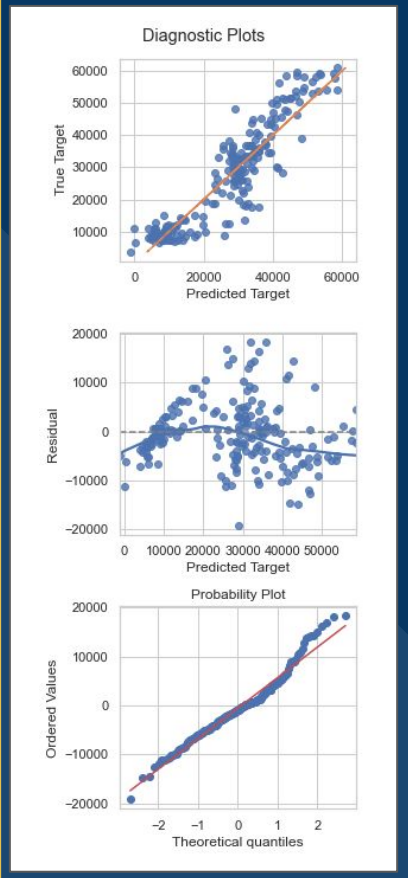
- Private schools generally become more expensive with high student scores, with a cap.
- Public schools offer a similar price point regardless of test scores





Relative Weights

- Taking care of students (sports, graduation rate, nice housing) translates to tuition growth.
- Public vs. Private and Admissions data is most significant
- LASSO dropped highly collinear features (population, crime)

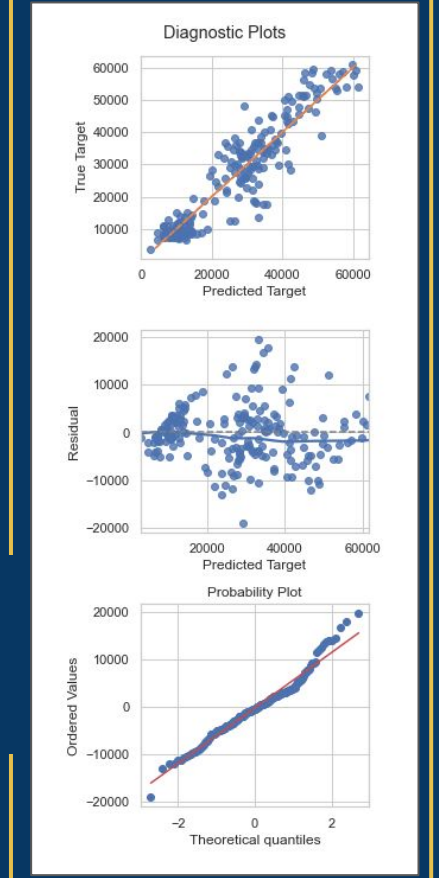


Before Scores vs. Public

Captures the relationship, with an undiagnosed curve, r^2 = 0.83 on Lasso Model

After Scores vs. Public

Model drops it's residual curve, r^2 = 0.87 on Lasso Model



1 Demographics and Population held apart

- Faculty Ratio vs. Population
- Gender vs. Race vs. Age

2 Location and Status matter

- Private vs. Public
- City vs. Country
- Test Scores

3 Schools need to take care of students

- VAWA (Violence Against Women Act) weighted more than other crime
- Sports Teams
- Expensive (Quality?) Room and Board

NOTABLE DEALS

Oberlin

Predicted: $46,500

Actual: :$58,554

BYU

Predicted: $40,500

Actual: $5,970

Wellesley

Predicted: $46,300

Actual: $58,448

Princeton

Predicted: $62, 800

Actual: $48,502

NEXT STEPS

Different Models/ Features/Data

Random Forest and Gradient Boost give insight into potential accuracy,, missing data could change results

Public vs. Private

Public vs. Private relationship deserves deeper analysis, possibly using historical data/separate modeling

Outcomes Vs. Tuition

Tuition is a model of market value, with MAE/MSE ~$5k; student outcomes may better capture intrinsic value

Thank You!