Capstone Project 2023
Financial Analysis Based on School Districts

Hongyu Dai
Cooperate with AllianceBernstein

Background

- Housing market is a key driver of the US economy.
- Housing prices have a significant impact on both consumers and investors.
- Prices of houses in neighborhoods with good school districts tend to be higher.

As a result, there is a strong relationship between school district quality and housing prices.

Overview

Difficulty

- Quantifying the relationship between school district quality and housing prices is difficult.

- Factors that affect housing prices are numerous and complex.

- Obtaining consistent school district data can be challenging.

- No readily available data

Contribution

- Developed pipeline for web scraping school district data and analyzing relationship with housing prices

  - Collect data on factors impacting school district quality

  - Use statistical techniques to analyze data

- Potential applications include identifying undervalued properties and informing education investment decisions
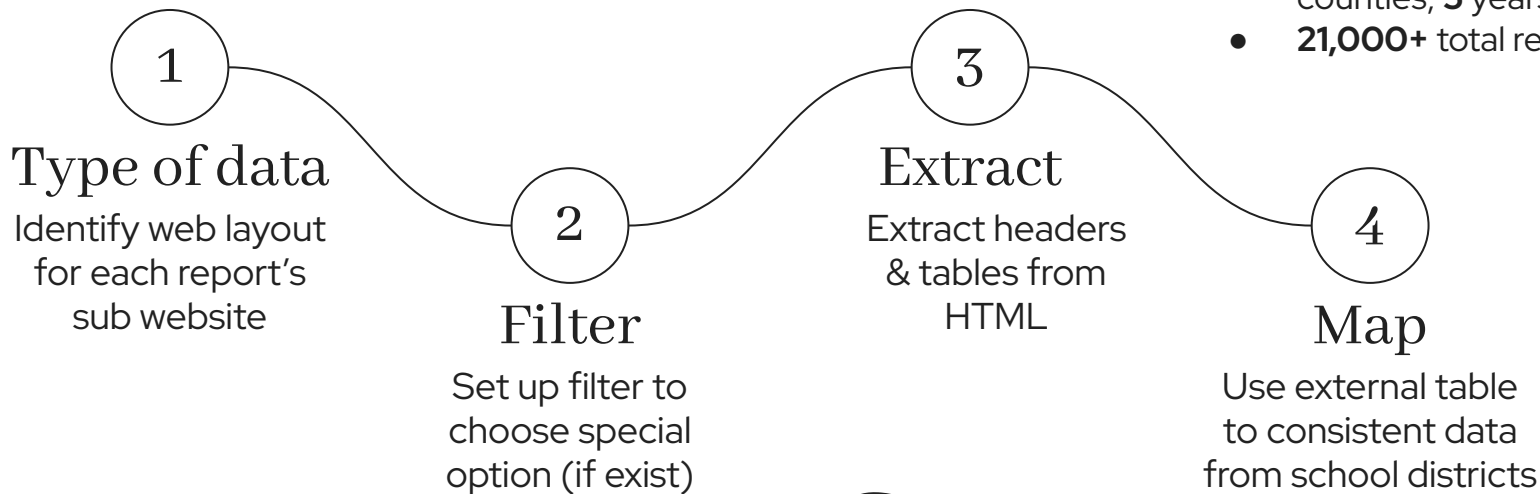
Data Collection - Web scraping

Summary

- Use Selenium & Beautiful Soup packages to simulate human actions
- Process collected HTML source code

Data Size

- **30+** hours of running time
- **20+** report types, **50+** counties, **5** years of data
- **21,000+** total reports

1

Type of data
Identify web layout for each report's sub website

2

Filter
Set up filter to choose special option (if exist)

3

Extract
Extract headers & tables from HTML

4

Map
Use external table to consistent data from school districts

Analysis

Objective

- Target variable: external housing price data
- Influencing factors: graduation rate from previous collected data

Data Processing

- Na value process
- Feature engineering: 4 years average, percentage change, lag change conversion
- Time node processing: selecting different time nodes to eliminate the possible impact of COVID-19

Exploratory Data Analysis

- Grouped histogram, scatter plot, and correlation matrix plot

Modeling

- Random forest regression model
- Linear regression model

Result

- After comparing different datasets, the linear model performed best on:
  - Data before covid-19
  - Target: the change rate of average house price per square foot in each school district over past 3 years
  - Coefficient of graduate rate change was over 0.62

- Graduation rate is a significant factor affecting housing prices.

- Results can provide valuable insights for policymakers and investors in making informed decisions about investments in housing and education.
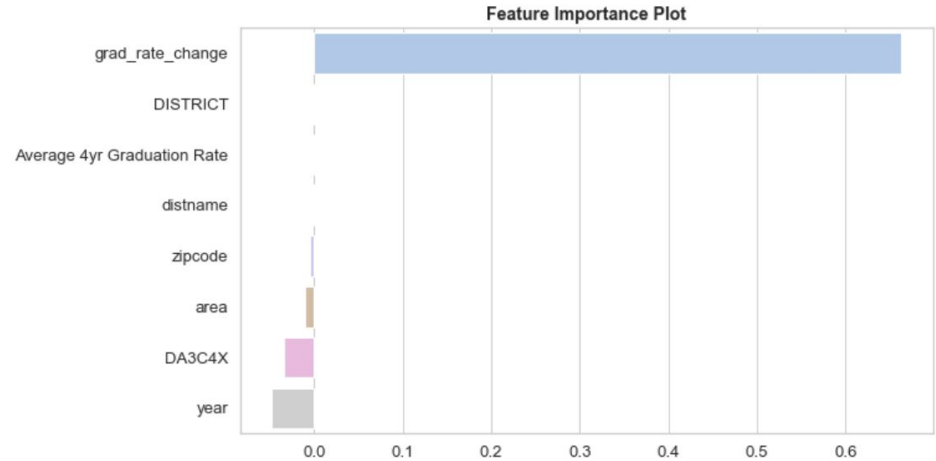
Feature Importance Plot
linear model

Thanks!

Do you have any questions?