Data Science Final Project

Brandon Feder, Atticus Wang

May 30, 2022

Introduction

Research question and hypothesis

Princeton has many diverse houses ranging in size, location, age, and many other factors. With this comes a huge range of house prices. But how do the characteristics of a house determine its price? Our research studies how factors such as the number of bedrooms, location, and the year built affect the price of the house. More precisely, we try to predict `soldPrice`, the price at which a house is sold, using the following possible predictors:

- nbhd (neighborhood)
- bed (number of bedrooms)
- fullBath (number of full baths)
- halfBath (number of half baths)
- style (style)
- age (yearSold minus yearBuilt)
- marketDays (days the house was on market)
- yearSold, daySold, monthSold (date at which the house was sold)

Based off our own experience, we predict that neighborhood/address and the year built have a significant impact on the price of the house.

Data description

Beatrice Bloom, a Princeton Residential Specialist, provides many great resources about the Princeton housing market including a table of houses sold in Princeton since 2011. This data can be found here. We intend to use this data to answer our question. The data is stored in `./data/pton-market-data.csv` in the Github repo.

Exploratory data analysis

Use `group_by` and `summarize` to find the top-10 styles and neighborhoods with the highest price:

Regression Analysis

Our final models are Lasso and PCR. Lasso has the lowest variance, while PCR has the highest R2.

Discussion and Limitations

Assumptions possibly not met, curse of dimensionality, did not incorporate census data

Table 1: Top ten styles and neighborhoods with highest meanSoldPrice

style	meanSoldPrice	nbhd	meanSoldPrice
Condo	159841.2	Palmer Square	132404.4
Bungalow	377301.5	Carnegie Lake Area	236507.7
Flat	409580.7	Griggs Farm	276560.9
End-Unit	411399.5	Borough	292345.7
Townhome	443747.8	Township	319306.4
Raised Ranch	445209.0	Washington Oaks	387626.0
Cottage	465925.0	Campbell Woods	390685.4
Ranch	471197.1	Town & Country	429419.1
Twin	487903.8	Queenston Common	435689.0
Cape	522943.6	Fieldwood Manors	454025.1

Conclusion

Additional Work

Introduce the other models we tried