

Data Science Final Project

Brandon Feder, Atticus Wang

May 30, 2022

1. 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.

For a simple exploratory data analysis, we used `group_by` and `summarize` to find the top-10 styles and neighborhoods with the highest price. The results are shown in Table 1.

2. Regression Analysis

Description of models

The final model we adopted was Lasso (short for “Least Absolute Shrinkage and Selection Operator”), a generalization of usual linear regression. We chose this model because among all models we tried (see Table 2), Lasso has the lowest variance of errors (an RMSE of roughly 0.453 million dollars), and a decent R-squared value of 0.522.

Lasso tries to enhance prediction accuracy and model interpretability by performing both variable selection (selecting which predictors to take into account) and shrinkage (shrinking coefficients of less important predictors). For usual linear models, we often try to minimize the sum of squares $\sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{i,j} \beta_j)^2$.

Table 1: Top ten styles and neighborhoods with highest meanPrice

| style | meanPrice | nbhd | meanPrice |
|--------------------|-----------|-------------------|-----------|
| Georgian | 2338700 | Pretty Brook Area | 1270498.3 |
| Transitional | 2188875 | Institute Area | 985201.3 |
| Manor | 1736780 | Western Section | 883958.9 |
| French | 1496095 | The Glen | 883113.5 |
| Mid-Century Modern | 1484571 | Hun Area | 871096.2 |
| Craftsman | 1385000 | Princeton Ridge | 772319.6 |
| Tudor | 1155930 | Battlefield Area | 762876.4 |
| Cape Cod | 1132179 | Riverside | 761638.1 |
| Farmhouse | 1105813 | Rosedale Area | 757956.3 |
| Normandy | 992742 | Ettl Farm | 752933.8 |

For Lasso, we impose a penalty for more complex models: what we minimize is the sum $\sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{i,j} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|$. Here, $\lambda \geq 0$ is a parameter that we can tune to best fit our scenario.

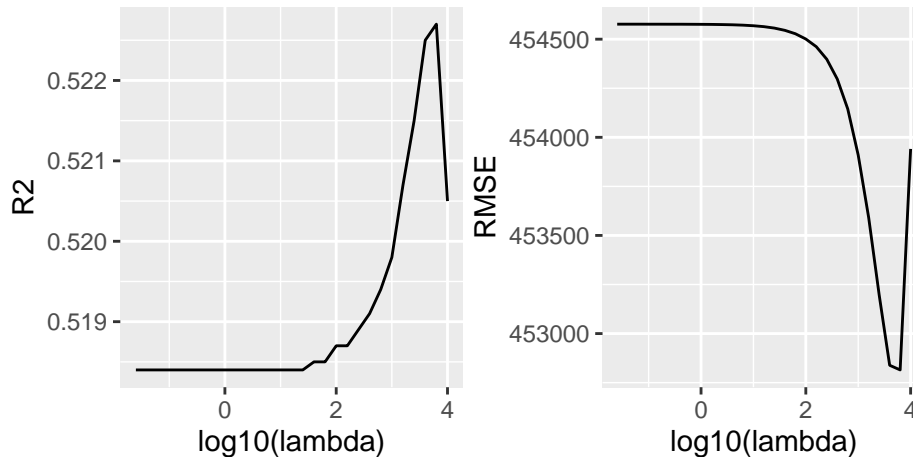
As usual, Lasso assumes that the response variable y follows a linear relation with the predictor variables $y = X\beta + \epsilon$, where ϵ is Normally distributed with mean 0. Also, since all predictor variables are assumed to be quantitative, we transformed categorical variables such as `nbhd` into dummy variables.

Table 2: Performance of all models we tried

| | R2 | RMSE |
|-------------|--------|----------|
| full lm | 0.5448 | 454579.1 |
| forward | 0.5449 | 454880.4 |
| backward | 0.5449 | 454880.4 |
| ridge | 0.5170 | 455475.4 |
| lasso | 0.5227 | 452814.7 |
| elastic net | 0.5192 | 454296.0 |
| pcr | 0.5996 | 507588.0 |

Model output

We calculated the average R-squared and RMSE values using tenfold cross validation. The following graphs plot R-squared and RMSE values against lambda, the parameter in the Lasso model.



As we can see, the best lambda occurs roughly at $10^{3.8} \approx 6309$. Using this lambda, we obtained the R-squared and RMSE values for Lasso in Table 2.

Interpretation of coefficients

Inspecting the coefficients chosen by the Lasso model, we found that among the total 61 variables (most of them dummy variables), only 41 of them have nonzero coefficients. The largest coefficients (in terms of absolute value) are `yearSold` (412690), followed by `fullBath` (232332) and `halfBath` (57121). This means that on average, house prices increase each year by around 0.4 million dollars, and each additional full bathroom increases house prices by around 0.23 million dollars.

Discussion and Limitations

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

Conclusion

Additional Work

Introduce the other models we tried