ECONOMIC MODELING OF CRAIGSLIST APARTMENT LISTINGS USING A “HEDONIC” GENERALIZED ADDITIVE MODEL

Grant B. Wiersum

Grant.Wiersum@Northwestern.edu

Abstract

Rental and real estate prices are among the most consequential factors of cost-of-living in America. The decision of how to price properties relies on a multitude of variables but very seldom are these decisions subject to critical analysis. A reverse question also exists. Under what conditions do landlords seek to renovate properties and increase rental prices? The rental pricing problem is therefore of great importance to landlords who wish to price their rental units competitively, for renters seeking maximum value and for public institutions seeking to understand patters of investment. In this paper, multiple methods are evaluated against the Chicago apartment listings as scraped from Craigslist.

Keywords: Chicago, real-estate, investment, pricing.

Introduction

Cities are dynamic, ever-changing landscapes. Over the course of urban history, numerous technologies have come into existence and faded out. Transitioning energy sources from wood and livestock to coal, then to petroleum and natural gas, and now to electricity each took significant investment. The ultimate societal reasons behind each transition center on cost, efficiency, and environmental impact (Osterbring, Mata, Thuvander and Wallbaum 2019). Tracking these changes, the heights of buildings in urban centers have grown in step with the availability of energy to supply to its inhabitants, but also notably, inversely to the costs of construction. (Ahlfeldt and McMillen 2017).

Housing prices remain unique among commonly traded goods and services, in that nearly no geographic competition exists. This feature is often compounded by high costs and low frequencies of transactions. Demonstrating the importance of location in urban centers, is the extreme segregation of central business districts which radiate outward into residential areas. (Lucas and Hansberg, 2002).

Apartment pricing in the residential real estate market remains stubbornly inefficient. With roughly half of US rentals owned by small, "mom and pop" investors, valuation errors are predictably commonplace. Anchoring effect has been demonstrated to play a significant role in price setting. Landlords who buy at a market's peak systematically charge 2-3% more and sit in inventory 6% longer than counterparts who purchased at the market's trough (Giacoletti and Parsons. 2022). This inefficiency is at least partially explained by the conventional wisdom of the "1% rule of real estate", which states rent should equal 1% of the purchase price. These peculiarities and inefficiencies underscore the need for a deeper understanding of price setting.

The formal approach to this problem is known as hedonic modeling which attempts to determine the extent to which each factor affects a property's price. The goal of this paper is to model the specific decision to renovate existing rental units through the development of such a model. By investigating the wealth of information publicly available on apartment listing sites, we can construct a model that accurately explains the bulk of the variation in price between rental units. By mining data available through Craigslist, a predictive model of investor action can be built, and a prescriptive model for investors in the residential real-estate space.

Literature Review

The problem of hedonic regression has been approached in numerous other fields. A method described by Rosen (1974) outlines methods since adopted by numerous institutions to built explanatory models of real estate pricing. Su et al. (2021) used this approach to good effect in a paper with similar goals, using web-based classifieds to investigate the role of landscape amenities on rental prices in China. The hedonic model attempts to decompose a property into its constituent attributes – square footage, number of bathrooms, bedrooms, etc. The advantage of this model is that it transparently accounts for the attributes of a property and is easily intelligible. Because of its ease of use and interpretation, the hedonic model has become industry standard for valuation in the American housing market. Hedonic models are often limited though, by poor generalizability.

To address the problem of generalizability, the generalized additive model (GAM) seeks to sacrifice comprehensibility in favor of better accuracy. Detailed by Mason and Qugley (1996), this model generates highly accurate estimates of price, but interpretability does suffer somewhat. Even so, this approach's strength lies in its ability to generate smooth functions from discontinuous datasets. Rather than depending on a nearest-neighbors method, a GAM directly calculates a gradient and provides a confidence interval.

Dataset Preparation

The data-gathering step was carried out using Selenium web-browser and Chromedriver for Linux. The scraping module first iterates over each page of listings, generating a series of URLs. The next module iterates over the listing URLs and generates a "Listing" object which stores the attributes of each unique listing. Attributes gathered are:

'price', 'beds', 'sqft', 'parking', 'baths', 'descript', 'adress', 'lat', 'lon', 'date', 'cats are OK - purrr', 'dogs are OK - wooof', 'air conditioning', 'furnished', 'w/d in unit', 'laundry on site', 'laundry in bldg', 'no laundry on site', 'no parking', 'street parking', 'off-street parking', 'detached garage'.

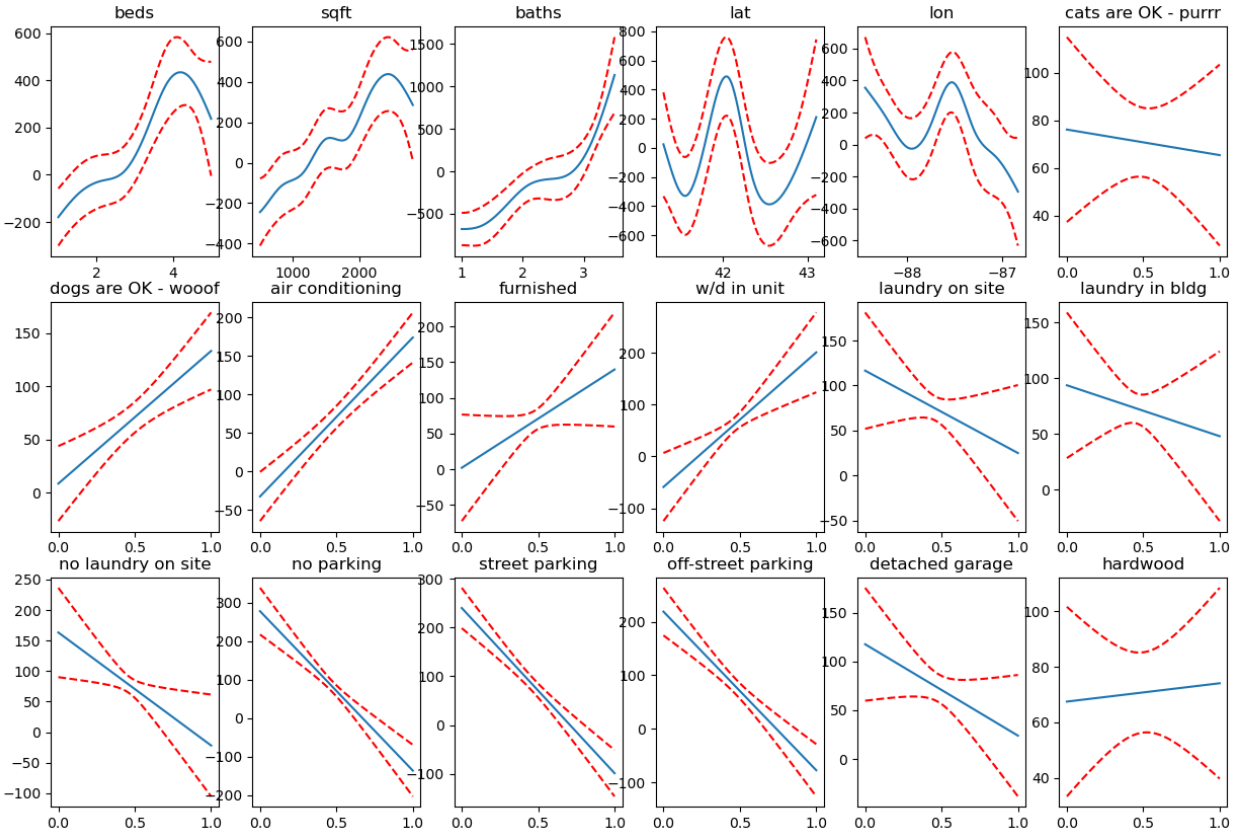
price	0.553906
w/d in unit	0.358435
air conditioning	0.256151
lat	0.223903
dogs are OK - wooof	0.202458
cats are OK - purrr	0.125219
stainless	0.109931
furnished	0.048016
hardwood	0.018486
lon	-0.022421
detached garage	-0.029638
laundry on site	-0.085805
off-street parking	-0.087408
no parking	-0.089571
street parking	-0.093387
laundry in bldg	-0.115613
beds	-0.200902
no laundry on site	-0.234182

Data was then stored in a Pandas dataframe for analysis. The variable “descript” is unique because it contains the entire free-text field provided by the listing entity. From this attribute, the fields “hardwood” and “stainless” were generated by simple string matching and one-hot encoded.

When the data were de-trended by using user-reported square-footage measurements, correlation values could be calculated. Of note, in-unit washer/dryer, air-conditioning and proximity to Chicago’s downtown correlated strongly with higher price-per-square-foot. Landlords reporting no laundry on site report an average price $1100/month lower than those reporting in-unit laundry. Rates of indicators of renovation, ‘stainless’ keyword and listed air-conditioning are also lower in laundry-absent properties. This provides support for the theory that these attributes can be largely considered together as indicators of property investment.

Still unclear though, is the actual hedonic value of adding any specific investment. To determine these marginal values in better detail, a GAM was trained on the dataset. By plotting the curves generated, we see the effects of each variable in isolation, along with their

respective 95% confidence intervals.



From this point, it becomes possible to provide guidance to maximize return on investment. The addition of air conditioning, for example, is valued at $75 to $200 per month. In-unit laundry, we see providing a benefit from $100 to over $300 per month. Updated kitchens (though not on the above figure), lend an additional $125 in value. As an example problem, assume a landlord would like to determine the additional value adding air conditioning would contribute to a rental unit. At time of writing, an 8% APR home-equity loan is typical for well-qualified borrowers. The EPA recommends 20,000 BTUs for the typical 1000 square-foot Chicago apartment. Prices for ductless split units vary widely and can range from as little as $3000 to upwards of $20,000. A washer/dryer set typically costs from $700 to $1800, and a stainless kitchen set (refrigerator, stove, dishwasher, and microwave) ranges from $2,300

to $11,000. This gives us a range of potential budgetary possibilities $6,000 to $32,800. The possible upside benefit is in the range of $225 to $700 per month or $2,700 to $8,400/year.

Through use of the confidence intervals provided by the GAM, we can simply add the mesh grids to generate predictive values for the expected change in rental incomes. The CI95 return on investment is found to range from $280 to $887, with a mean of $498 per month, or $5,976 annually, and $59,760 over the lifetime of the loan.

Discussion

The key drawback of this study is that while it performs well predicting the mean asking value of a property, it fails to take time-on-market into account. The monthly nature of most rental agreements implies that there is an optimal time-on-market for listings looking to maximize revenue. Listing a property at too low of a price may secure a tenant quickly, but leave money on the table. Listing a property at higher prices corresponds to longer times on-market, but regardless of which scenario, a move-in date will typically remain the same. Landlords should therefore ask exactly as much as they can without excessively risking the loss of a month's rent. To further investigate this, 90 days or more of rental listing data could be used to track the rate of posting renewal (Craigslist posts are renewed weekly) and attempt to learn from their correlations with this model's predictions.

Further enriching this model is also possible. With advances in computer vision, it may be possible to train a neural network to perform regression analysis of posting images and thereby provide guidance to an interpretable model such as a GAM.

References

Giacoletti, M., and Parsons, C. A. 2022 *Peak-Bust rental spreads.* Journal of Financial Economics

Goodman, L and Mayer, C. 2018. *Homeownership and the American Dream*. The Journal of Economic Perspectives.

Lucas, R. E., & Rossi-Hansberg, E. 2002. *On the Internal Structure of Cities*. Econometrica

Mason, C. and Quigley, J.M. 1996 *Non-Parametric Hedonic Housing Prices* Housing Studies.

Osterbring, M., Mata, E., Thuvander, L., Wallbaum, H. 2019. *Explorative Life-cycle assessment of renovating existing urban housing-stocks.* Building and Environment.

Rosen, S. 1974. *Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.* Journal of Political Economy.

Appendix

Linear GAM output summary:

```
LinearGAM
============================================= ==============================================
=========
Distribution:                        NormalDist Effective DoF:                     38.0656
Link Function:                     IdentityLink Log Likelihood:                -11505.5946
Number of Samples:                         890 AIC:                             23089.3204
                                               AICc:                            23093.0035
                                               GCV:                            177810.4216
                                               Scale:                          164175.7792
                                               Pseudo R-Squared:                    0.7356
============================================================================================
=========
Feature Function                  Lambda               Rank         EDoF         P > x        Sig
. Code
================================= ==================== ============ ============ ============ ===
=========
s(0)                              [0.6]                10           5.0          1.65e-07     ***
s(1)                              [0.6]                10           5.7          1.56e-04     ***
s(2)                              [0.6]                10           4.3          1.11e-16     ***
s(3)                              [0.6]                10           4.7          1.11e-16     ***
s(4)                              [0.6]                10           4.2          3.77e-14     ***
s(5)                              [0.6]                10           1.1          8.95e-01
s(6)                              [0.6]                10           1.1          9.74e-03     **
s(7)                              [0.6]                10           1.1          1.25e-09     ***
s(8)                              [0.6]                10           1.1          2.85e-01
s(9)                              [0.6]                10           1.0          4.32e-03     **
s(10)                             [0.6]                10           1.0          4.93e-01
s(11)                             [0.6]                10           0.9          8.90e-01
s(12)                             [0.6]                10           0.9          1.79e-01
s(13)                             [0.6]                10           0.9          1.09e-09     ***
s(14)                             [0.6]                10           0.9          4.63e-11     ***
s(15)                             [0.6]                10           0.9          6.53e-09     ***
s(16)                             [0.6]                10           0.8          4.24e-01
s(17)                             [0.6]                10           0.9          6.49e-01
s(18)                             [0.6]                10           0.8          7.18e-04     ***
s(19)                             [0.6]                10           0.8          2.65e-01
intercept                                              1            0.0          1.11e-16     ***
============================================================================================
=========
Significance codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

WARNING: Fitting splines and a linear function to a feature introduces a model identifiability pr
oblem which can cause p-values to appear significant when they are not.

WARNING: p-values calculated in this manner behave correctly for un-penalized models or models wi
th known smoothing parameters, but when smoothing parameters have been estimated, the p-values ar
e typically lower than they should be, meaning that the tests reject the null too readily.
```