Optimal Home Pricing Prediction

Wesley Stevens
Jacob Barazoto

April 9, 2019

1 Abstract

Real estate is one of the largest markets in the economy. Previously, brokers have identified optimal house prices on a case-by-case basis. Using machine learning techniques we propose a model that predicts the selling price of a home in the Western United States with 92% accuracy.

2 Problem Statement and Motivation

The housing market is a large and profitable industry; many people make a fortune by flipping houses. We want to automate the process of finding a marketable house by building a predictive model to determine its optimal selling price. In doing so, we will also identify what house features are most important in determining the price. Although a similar service is currently provided by some real estate websites, only an estimate of the selling price is listed. Their processes of identifying important features are not publicly known nor is the process of getting this estimate. We will build our own model to obtain these important features by gathering, engineering, and applying machine learning algorithms to current real estate data.

3 Ethics

Providing the optimal price of a house and determining the most important features in that process seems to have no ethical problems in and of itself. Any potentially negative impacts our research could have has already been introduced by other pricing services. We merely offer detailed information on how those prices are determined. In gathering our data, there were website scraping restrictions regarding what information we could use, but we followed these rules strictly to stay within bounds of ethical practices.

4 Data

After researching real estate websites that contained current and reliable data on recently sold houses, we chose to gather data from *estately.com*. Their minor scraping restrictions allowed us to legally obtain all the data we needed to build our models. Since they offer real-time updates on houses in the real estate market and relied on licensed real estate agents for that data, we were confident that it was accurate.

We implemented a depth first search algorithm to traverse the listings of each house in each city in California, Arizona, Colorado, and Nevada. We paused for each web page transaction to avoid slowing down the website's server.

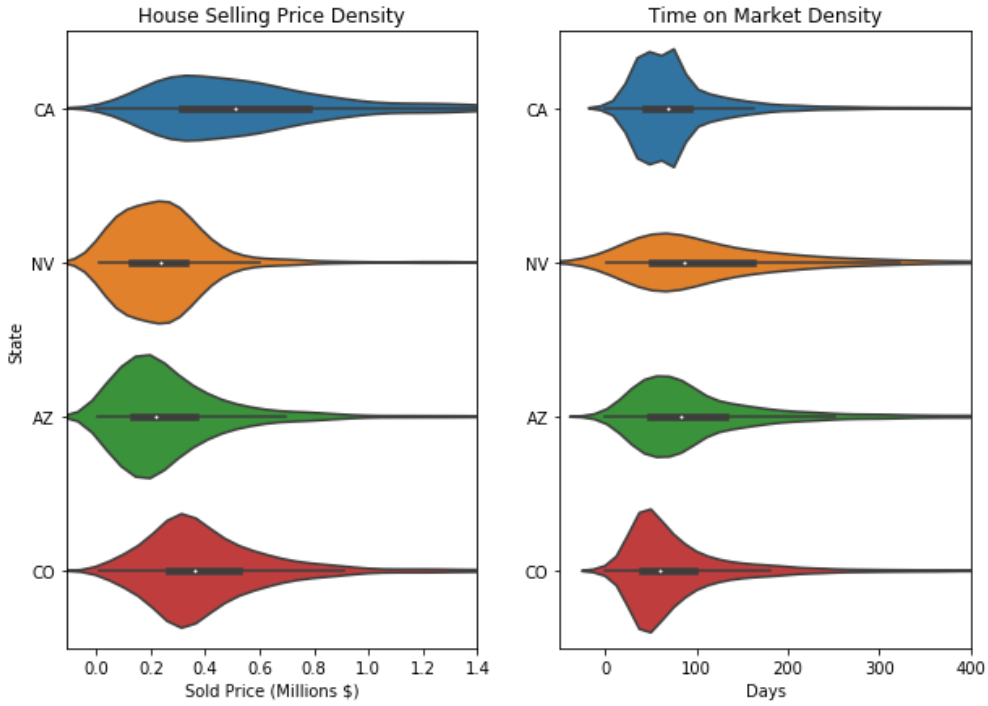
Our data initially contained information on ~100,000 listings in these states, with 22 features per listing. After cleaning and feature engineering what we gathered, 73,000 total listings remained with 37 features per listing. We determined it was necessary to drop certain features and listings in order to minimize bias and improve accuracy. Many of the added features were just a one-hot encoding.

Below we provide a sample listing from our data. The house size is in square feet, the lot size is in acres, the sold price is in millions, and the elementary school rating is out of ten. We also one-hot encoded by state, month it was posted, month it was sold, and if it was a foreclosure sale.

Bedrooms	6.00
Bathrooms	3.00
HouseSize(sqft)	1956.00
LotSize(acre)	0.09
YearBuilt	1935.00
Stories	1.00
SoldPrice	1.20
UtilityCosts	190.00
ElementarySchoolRating	10.00
DaysOnMarket	165.00
Foreclosed_True	0.00
State_CA	1.00
State_CO	0.00
State_NV	0.00
Posted_Feb	0.00
Posted_Mar	0.00
Posted_Apr	0.00
Posted_May	0.00
Posted_June	1.00

Posted_July	0.000000
Posted_Aug	0.000000
Posted_Sept	0.000000
Posted_Oct	0.000000
Posted_Nov	0.000000
Posted_Dec	0.000000
Sold_Feb	0.000000
Sold_Mar	0.000000
Sold_Apr	0.000000
Sold_May	0.000000
Sold_June	0.000000
Sold_July	0.000000
Sold_Aug	0.000000
Sold_Sept	0.000000
Sold_Oct	0.000000
Sold_Nov	0.000000
Sold_Dec	1.000000
Marketability	0.005492

To familiarize ourselves with our data, we visualized various aspects. In one such visualization, we plotted the selling price distribution of each home in our data set according to the state the house was in. As seen in the plots below, houses in California are generally priced higher than houses in other states.



5 Methods

After talking with real estate brokers, we discovered that once a listing is posted, it should be getting views after the first week and offers within a month. The average time to close a deal was 2.5 months after the house was originally posted. So we included a indicator to quantify this behavior's effect on the house's selling price. This indicator is a normalized error function with a punishing coefficient. The punishing coefficient was made based on how long the house was on the market. The longer the house is on the market, the more it punishes. It takes as input the time a house is on the market and punishes the model the longer the house is listed.

We constructed training and testing sets using an 80/20 randomized cross validation split. This allows us to train and test our models while avoiding introducing potential bias. Since our data was continuous, it made sense to build our models using regression algorithms, and not classifiers. To construct the best model, we compared the accuracy of following algorithms: Support Vector Machine regression, K-Nearest Neighbors, Kernel Ridge regression, Partial Least Squares regression, an Extreme Gradient Boosting (XGBoost) regressor, and a Gradient Boosting Trees regressor. We also tuned several hyper parameters via a cross validation grid search to boost accuracy. Parameters included: γ, α, learning rate, and λ.

6 Results

Support Vector Machine Regression

SVM regression is a logical first choice since it is effective in high dimensional spaces while offering efficient computation. However, despite fine-tuning hyperparameters, we only achieved 6% accuracy. We were unable to identify the cause of such low accuracy.

K-Nearest Neighbors

K-Nearest Neighbors offers cluster-based learning on continuous labels. Other benefits include low temporal complexity on high-dimensional spaces. We achieved a 50% accuracy rate with this algorithm, which was worse than expected, but better than SVM regression.

Kernel Ridge Regression

Kernel Ridge regression seemed like a good choice since it projects high-dimensional data to smaller sub-spaces allowing for more efficient computation. We achieved a 16% accuracy score which was disappointing. This was probably caused by data loss through projection.

Partial Least Squares

We hoped to find a linear relation between the data and the selling price. Partial Least Squares regression finds linear relations between multivariate data sets which made it a good choice for determining if this is the case. However, we achieved an R^2 score of -0.685, which means the relation is not linear.

XGBoost Regression

XGBoost provides a parallel tree boosting of random forests. It benefits from deterministic model building while having a loss function that severely pushes over-fitting. Fitting our data with fine-tuned hyper parameters gave us a score of 88.97% accuracy. This is both reliably accurate and well-fit since each decision tree uniquely predicts each feature while not over-fitting or overlapping. Thus, we believe that this algorithm offers a potentially accurate predictive model for optimal pricing.

Gradient Boosting Trees Regression

Gradient Boosting is very similar to XGBoost. It differs mostly in tree penalisation and leaf node size determination. However, the core concepts remain the same: it uses numerous decision trees to avoid over-fitting. We obtained an accuracy score of 91.8% in this model. Seeing as this model is very robust to over-fitting, it would also be a good choice for a predictive model.

7 Analysis

The 92% accuracy rating is very good. A higher accuracy rating might be achieved using more features per listing, or introducing more listings from other parts of the country. The accuracy might also increase if we improve our model for punishing market

time. The current model is a basic yet a highly functional representation of indicting optimal selling price. Applying neural networks or other regression algorithms may also improve the model's fit, but may not provide significantly better results in exchange for the temporal efficiency of boosted trees.

Using the Gradient Boosting Trees regressor and a 95% confidence interval, the most important features to determine optimal pricing were: the number of bathrooms, size of the house, year it was built, and time on market. Most of these are features we initially expected to influence selling price, but we are surprised to not see features such as: the number of bedrooms, location, and lot size. Although they certainly do have an impact on its marketability, our model suggests that they have little impact upon its selling price.

8 Conclusion

In the end, we were successful in creating a model that accurately predicts the optimal selling price of a house in the Western United States. The results offer insights on what features most prominently affect selling price while demonstrating that data offers information that human intuition may deem insignificant.

So given a choice between adding a bedroom or a bathroom to your home, choose the bathroom.

9 Bibliography

1. "Homes for Sale, MLS-Based Real Estate." Estately, www.estately.com/.
2. "Homes for Sale, MLS-Based Real Estate." Estately, www.estately.com/robots.txt.
3. "Machine Learning in Python." Scikit, scikit-learn.org/stable/.
4. "Zestimate" Zillow, https://www.zillow.com/.