Semester Project

Joo-Wang John Lee*, Andrew Patterson*, Matthew Schlegel*

Abstract

This paper starts a conversation in using data mining and more principled statistical techniques to evaluate the prices of houses. Using data provided from the Iowa town of Ames several techniques are explored to create a model for predicting housing prices. The best model used was a logistic regression model using a PCA feature reduction. Over several runs the percentage average mean error using 242 PCA features was 9.699%.

Keywords

House Price Prediction — Regression — PCA

[1] *Computer Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
***Corresponding authors**: jl216, andnpatt, mkschleg@indiana.edu

Contents

Introduction

The housing market is one of the most wide-reaching markets in the United States. Whether it be buying a condo in the middle of New York City or a small cottage in the Appalachian mountains, many people are in need of a reliable source of information about housing conditions and costs. Valuation of a home based on different factors such as geography, space, architecture, etc. is a fluid process that must hone and cater to market trends and the interests of home buyers. In the real estate market, trained statisticians deliberate on various features of a home such as geography, space, and others, and give prognostications on the value of a home based on weighing these different factors.

While the housing market is pervasive and important to the majority of Americans, it is fraught with corruption [1]. Pricing manipulation is a major concern for those looking to buy houses and can cost the average consumer hundreds of thousands of dollars. Sadly in the past, the over-valuation of houses and the easy-to-get loans available from many banks has caused not only the housing market but also the American economy to suffer through a major recession.

The future of house valuation must be through non-partisan non-human measures, focusing on several key points and features of a given house and producing a fair valuation based on the prices of houses sold before it. A question arises in that houses are evaluated on many features, and the data provided by the housing commission is often incomplete and hard to manage. The weighing of factors to predict prices can be done more systematically using data mining techniques. Taking the housing data in Ames, Iowa provided by Kaggle based on 2930 observations in the years 2006-2010, we analyze it using various regression techniques to weigh the different features of homes to determine and predict the most fitting price for said home.

This paper presents the following: (1) Highlights several methods for evaluating a houses cost and several methods for dealing with the data management and cleaning. (2) Describes more fully the background of the problem faced in this paper. (3) Contains a deeper discussion about the methods and approaches used for cleaning/modifying the data and for creating a model to predict the cost of a house. (4) Contains the design and results of the experiments run with the methods. (5) Contains conclusions and an overview of the work contained in this paper as well as future work needed to further the accuracy of the model.

1. Background

The problem faced in this paper is in the prediction of house prices for a small town called Ames located in Iowa. Due to the corruption and manipulation of the housing market, a model for accurate and fair price predictions is needed. Several techniques have been created and explored to predict or model house prices. The most commonly used method is the adjustment-grid method [1] described by Gau and Wang and was first provided with an analytical foundation in 1983 by

Colwell, Cannady, and Wu [2]. Other grid methods have been created combining the adjustment-grid method with popular regression techniques searching for a more robust and sound approach to weighting property features [3].

The problem is simply explained in terms of linear regression [4] where some function of the given features and the targeted value and produce a weight vector such that

$$\begin{aligned} x &= X(housesize, lotsize, streettype, etc...) && (1)\\ y &= Y(houseprice) && (2)\\ \hat{\theta} &= (x^T x)^{-1} x^T y && (3)\\ y_{pred} &= \hat{\theta}^T x. && (4) \end{aligned}$$

With this model, predictions can be made to approximate the price of a house based on the features provided by the house appraisers. Other considerations need to be made to the data to ensure the model has the best chance at producing accurate predictions. Many appraisers take heavy consideration into the location of the house, and also base housing prices on the general health of the market.

2. Algorithm and Methodology

Several methods were used for regression and for data cleaning/modification. Below we list the successful methods used in the experimental section, and the data manipulation techniques used to clean and conform the data into a reasonable set to train the predictive models.

Data

The housing data was relatively manageable in size, containing only 1460 samples. The major problem with this data set was the set of features given for prediction. Many of these features were words and classes which have a less obvious impact on the price of the house compared to values such as the size of the house or lot. What needed to be done was a reduction in the number and complexity of the given features. The major features that seemed to be the most useful from our intuition are the LotArea, LandSlope, Neighborhood, BldgType, OverallQual, OverallCond, YearBuilt, YearRemodAdd, BsmtQual, BsmtCond, Functional, YearSold, SaleType, CondSale. These features appeared to encompass a broad sense of the houses quality, size, and location which all could be useful to predicting the price. One other major concern was overfitting to the training data. Because of the large space of features available, it was more efficient to create a representation that was as simple as possible.

Data Cleaning and Modification

The housing data provided included a number of missing data points needing replaced. One such column was the Lot Frontage in which the missing values were replaced with the average ratio of lot frontage to lot size multiplied by the lot size of the given property. There were also missing values in several of the class columns. In these instances, three options were taken: deleted the column all together if there were many missing values, added a new class of NA for columns that it made sense such as Alley (assuming NA means no alley access), and replaced the missing data selecting from a distribution calculated from the class probabilities. The data was heavily class dependent, with many categories having a fixed number of classes. To overcome this issue, the data was coded such that each class received its own binary feature.

Several features were modified to values that would be more representative of the data. Values such as the year built and year remodeled were subtracted from the year the house was sold to create a better sense of the age of the house when sold. The year the garage was built was removed as there were several missing data points. The MasVnrArea was modified such that the NA values were replaced with 0, as we assume a missing value here means it didn't exist at the property.

Clustering

Location and other general factors were taken into heavy consideration during the valuation of a house. This included the neighborhood's condition and surrounding houses. To account for this, one method of implementation was clustering the data based on location and then performing regression on the clustered data. With this, there was a neighborhood by neighborhood model from which to predict. The clustering algorithm used was K-means with 25 centers representing the 25 different neighborhoods. Along with clustering, using just the neighborhood data, clusters were made using all the data as well.

Regression

Two regression models were used in this paper, linear and logistic. Linear regression is as stated above where we want some weight vector such that

$$\begin{aligned} x &= X(housesize, lotsize, streettype, etc...) && (5)\\ y_{actual} &= Y(houseprice) && (6)\\ \hat{\theta} &= (x^T x)^{-1} x^T (y_{actual} - \varepsilon) && (7)\\ y_{pred} &= \hat{\theta}^T x + \varepsilon. && (8) \end{aligned}$$

Where we want to minimize the error in between y_{pred} and y_{actual}. Logistic regression is similar where

$$log(\frac{y}{1-y}) = \theta^T x + \varepsilon. \quad (9)$$

Feature Reduction

The key method to handle the overwhelming amount of features in the data set was Principal Component Analysis (PCA). By mean centering and distinguishing principal components based on degrees of covariance and correlation, the number of features to analyze were reduced. Attributes that have similar correlations are grouped as new features based on their covariance matrices. The output would be helpful in better visualizing the data points that have more relevance in making predictions on housing prices.

3. Experiments and Results

Each experiment was run 20 times with a randomly sampled testing and training set. There were 1200 training points and 260 left for testing. The experiments ran on a mid-2015 Macbook Pro 15” with 2.5 GHz Intel Core i7 and 16 GB of 1600 MHz DDR3 RAM running macOS 10.12.2. The experiments were implemented in R using the following libraries: MASS, klaR, caret, e1071, glmnet, aod, ggplot2, Rcpp, and elasticnet. The sale price values were transformed by calculating the percentage of each value over 1 million in the set. The predicted value was then multiplied by 1 million and the error was calculated.

Model Type	Train Error	Test Error
Linear Not Sparse	0.112	0.127
Linear Sparse	0.076	0.120
Logistic	0.069	8.14
Elastic Net	10.43	11.57
PCA + Logistic 82	0.08	0.099
PCA + Logistic 242	0.04	0.075

From the above table PCA using 242 PCA features and logistic regression does the best. The below figure shows the plot of error with the number of PCA features used.

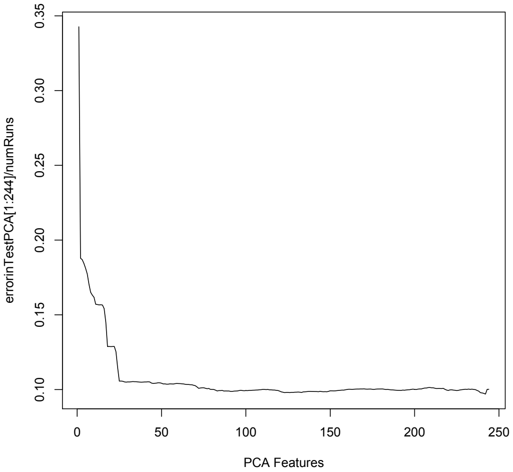

Figure 1. PCA Features

The experiments using clustering did not perform well and got error ranging from 100% to 3500%. I believe this occurred due to the lack of training samples the clusters were comprised of. Many of the blocks only had a few points while others contained a vast majority. With more data and possibly a more carefully designed implementation for clustering this may have performed better.

4. Summary and Conclusions

While this is the first step in discovering methods for predicting the sales price of houses, the above results shows promise in using data mining and machine learning techniques for the analysis of housing data. A more reliable model using a more robust data set could be produced, but the lack of data in the set limits the amount that can be learned and the hinders its ability to be generalized.

Future work should go into using better methods for regression and prediction, creating and training with a larger data set to produce a more reliable and general model, and more complex feature selection methods. Also using a larger dataset the clustering method should be re-approached and experimented with.

Acknowledgments

We thank Indiana University, especially Prof. Dalkilic, for their gracious and unwavering commitment to students, and the Lindley Hall staff for the amazing facilities (especially the coffee machine). Finally, we must thank the knowledgable and helpful AIs, and the omnipotent and omniscient Google.

References

[1] George Lentz and Ko Wang. Residential appraisal and the lending process: A survey of issues. *Journal of Real Estate Research*, 2009.

[2] Peter F. Colwell, Roger E. Cannaday, and Chunchi Wu. The analytical foundations of adjustment grid methods. *Real Estate Economics*, 11(1):11–29, 1983.

[3] George W Gau, Tsong-Yue Lai, and Ko Wang. Optimal comparable selection and weighting in real property valuation: An extension. *Real Estate Economics*, 20(1):107–123, 1992.

[4] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2015.