Predicting House Prices based on Descriptive Real-Estate Data

Eric Murphy '19, UIN# *****726

Abstract

This final project aims to delineate and systematically identify the key trends most responsible for predicting the price of real estate with respect to various estate features.

The Ames Iowa data set describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a sizeable number of explanatory variables, of which 23 are categorical variables, 23 are ordered variables, 14 are discretized, and 20 are continuous.

Upon data cleaning and some imputation of extraneous variables, 80 variables remained that were directly related to the real estate in question that focus on the quality and quantity of many physical attributes of the property.

Motivations for Analysis

Personally, I found interest in this dataset because I am or will soon be actively interested in purchasing a home someday; I found that the analysis would prove to be useful to me as a current student for learning as well as a (hopefully) salaried data practitioner one day that can afford one. As it would be, most of the variables are the kind of information that the average home buyer would want to know about a potential property, that answer **direct** and *indirect* looming questions such as:

1. **When was it built** / *is it too old to maintain long-term*?
2. **How big is the lot** / *is it too small relative to the price*? (e.g. coastal and metropolitan real estate is pricier for less room)
3. **How many square feet of living space is in the dwelling** / *where's that space used most or least*?
4. **How many full and half bathrooms are there** / *will my housemates have to share one, am I okay with that, if we have guests will there be a dedicated guest bathroom?*

Intuitively, these are useful variables that influence our natural decision-making process. In aggregating information we value highest, we naturally decide if the features justify the price.

Description

The 14 discrete variables often quantify the number of items occurring in the house, such as number of kitchens, bedrooms, and full and half-bathrooms, and above ground living areas of the home. The 20 continuous variables pertain to various size dimensions for each observation, usually in square-footage. In addition to the average lot size and total square footage found on most common home listings, other more specific variables are quantified in the data set, such as area measurements on certain rooms such as basements. The primary living area, and even patio space is reduced into individual categories based on quality and type. Remodeling dates are also recorded, which is practical information in house hunting. They range from two to 28 classes with the least described being STREET (gravel or paved) and the often-described being NEIGHBORHOOD. The nominal variables identify various types of dwellings, garages, materials, and environmental conditions while the ordinal variables qualitatively rate various items within the property.

Questions

The genesis for most questions for this project have their roots in human's ability to intuitively derive value from multivariate analysis in day to day life. As such, I primarily explore solutions to four questions:

1. Which variables explain most of the variance and fit of the predictions? Do any such variables make sense?
2. In doing our analysis, should we exclude highly suspect outliers. If the data is skewed, should we transform it or leave the distribution as-is? Is there regression towards the mean?
3. Which algorithm predicts best? I will explore LASSO, Principal Component Regression, Ridge Regression, ElasticNet Regression mixed models, Extreme Gradient Boosting (XGBoost), and Random Forest methods. Do any methods translate easily to natural intuitive processes?
4. Among discovered correlations, which ones are naturally unintuitive? Does this analysis provide scalable insight?