Project Overview

My project is to utilize machine learning to predict how much to negotiate on a house you're looking to purchase. The idea came to me when a friend of mine asked me help her come up with a price for a house she wanted to bid on. Many sites such as Zillow already provide an estimate of a house is worth based on their own proprietary formula. However, these estimates can be sustainably different than the price the seller wants to sell at.

Problem Statement

What makes my project different is I want to attempt to predict the difference between the listing price and the sale price rather than the actual value of the house. The listing price is the price the seller is advertising to sell their home for. The sale price is what the home actually sells for after negotiating with the buyer.

Metrics

The machine learning algorithms explored will be evaluated against one another using r^2. The coefficient of determination or r^2 tells how much of my variation in my output is explained by the change in my inputs. This metric is used for regression rather than classification since the values we are trying to predict are continuous rather than discrete. It is not affected by the size of the dataset unlike sum of squared errors.

Data Exploration

The data set was compiled and prepared by my friend's realtor. The data contains recent sales of homes within a 5 mile radius of the home she was looking to purchase. It came directly from the multiple servicing list service (MLS). The MLS databases allows real estate brokers who share their listing with one another for the purpose of locating ready, willing, and able buyers more efficiently. This database is only accessible by professional real agents.

PID – unique number identifying each real estate transaction

Subdivision – the subdivision the home is located in

 # Bedrooms – number of bedrooms in the home

 # Baths – number of bathrooms in the home

 # Rooms – total number of rooms in the home

Fin SF – total size of the house in square feet

List Price – the price the seller wants to sell the home for

Sales Price – the price the home actually sells for

Days On Markets – number of days the home has listed for sale on the market before being sold

Data Cleaning

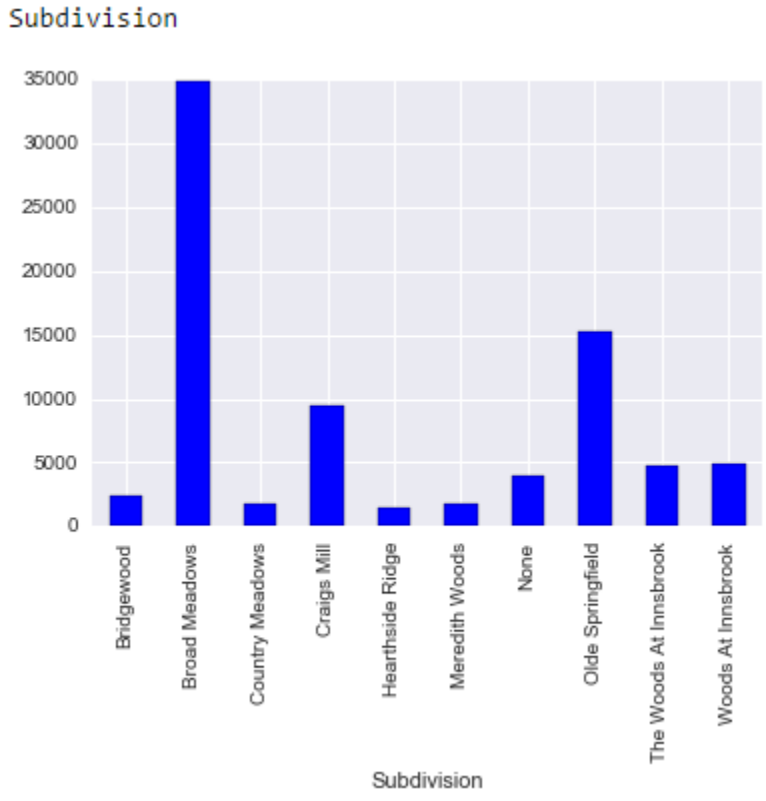
The data did not contain any null values that need to be removed. However, some columns had either negative or zero values which needed to be removed. Data points where # Rooms was less the sum of # of Bedrooms and # Baths were removed as well. Categorical data such as Subdivision were transformed into Boolean dummy variables. Target is the thing we're interested in predicting. It is computed by taking the difference between the List Price and Sales Price.

Numerical Feature Data Exploration

- # Baths has a correlation of 0.025 with the Target
- # Bedrooms has a correlation of 0.123 with the Target
- # Rooms has a correlation of 0.168 with the Target
- Days On Market has a correlation of 0.223 with the Target
- Fin SF has a correlation of 0.186 with the Target
- List Price has a correlation of 0.272 with the Target

Most numerical features had very little to no correlation with the target. The features with the highest correlations were List Price and Days on Market.

Categorical Feature Data Exploration



Subdivision

- **Bridgewood 2366.785714**
- **Broad Meadows 34900.000000**
- **Country Meadows 1715.000000**
- **Craigs Mill 9533.333333**
- **Hearthside Ridge 1450.000000**
- **Meredith Woods 1828.841270**
- **None 4000.000000**
- **Olde Springfield 15285.714286**
- **The Woods At Innsbrook 4743.043478**
- **Woods At Innsbrook 4873.194444**

The **Subdivision** appears to be more much promising Categorical Feature. It appear homes in located in Broad Meadows or Olde Springfield were significantly discounted.

Prediction

I constructed a nearest neighbor's regression to make predictions. Grid search was used to find the optimal hyper parameters were 19 for number of neighbors using uniformed weights. Unfortunately, even the model with the best hyper parameter had a negative r^2 which means the model failed to learn from the data.