Predicting Housing Prices in Ames, IA: GBM, PCR, and Linear Regression

Arina Voronina

MASTER OF SCIENCE IN Applied Economics

BOSTON COLLEGE
WOODS COLLEGE OF ADVANCING STUDIES

Introduction

What determines the price of a house? The neighborhood? The size of the lot? The number of bedrooms? The Ames housing dataset helps us explore the relationship between a house's various features and its sale price. With 2919 houses and 79 features for each house, this data is also a veritable playground for exploring data cleaning, statistical computation, and machine learning techniques that work around the quirks and challenges of a dataset.

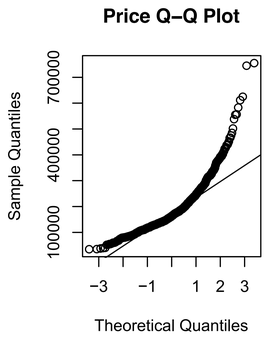
For this project, I compared the prediction capabilities of Gradient Boosting, Principal Component Regression, and Linear Regression. The goal was to achieve the highest accuracy possible in predicting housing prices in Ames, Iowa. With the data cleaning techniques and engineered features that I implemented, the Gradient Boosting model yielded the most accurate prediction.

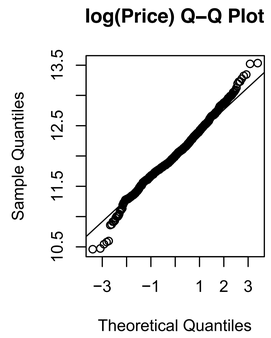
Methods

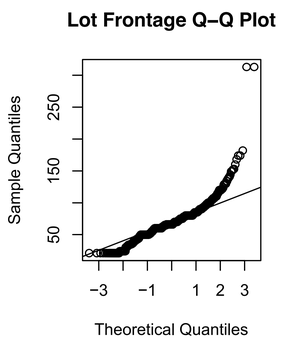
The Ames housing dataset contains 2919 observations from 2006 to 2010, with 1460 in the training data set and 1459 in the test dataset. Each house has 79 variables.

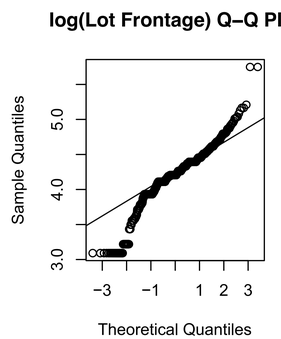
I started with basic data exploration and data cleaning. Missing Values in this data were most commonly due to a feature not being present in the home, like a pool. In the houses where the feature in question wasn't present, I filled the *NA*s of categorical variables with "None" and filled the numeric variables with 0. For other cases of missingness, I imputed a median on numeric variables or a mode on categorical variables. For some cases the value could be estimated using a reasonably related variable: features like Garage or Basement have several accompanying variables that can be used for cross-referencing. I then identified numeric vs categorical variables, taking care to correct misclassifications. Ordinal variables that indicate a rank, such as variables indicating quality from "Poor" to "Excellent", were assigned a numeric weight.

For feature engineering, I started by performing a log transformation on our target variable Sale Price in order to have a normally distributed response variable. I also created a new variable called House Age to capture the effect of a house's age on price.









QQ Plot: Sale Price (Top) and Lot Frontage (Bottom) before and after Log +1 transformation

The first thing I noted during my inspection of the variables was their skewness. 35 variables showed a skewness over |0.5|. Log-transforming skewed variables did not yield accurate results. This can be seen in one such variable, Lot Frontage, in the QQ plot above. Importantly, our response variable Sales Price became normal.

Another important facet of this dataset is multiple variables that describe the same feature, such as the variables that describe the house's basement, or all the numeric variables that detail a house's size.

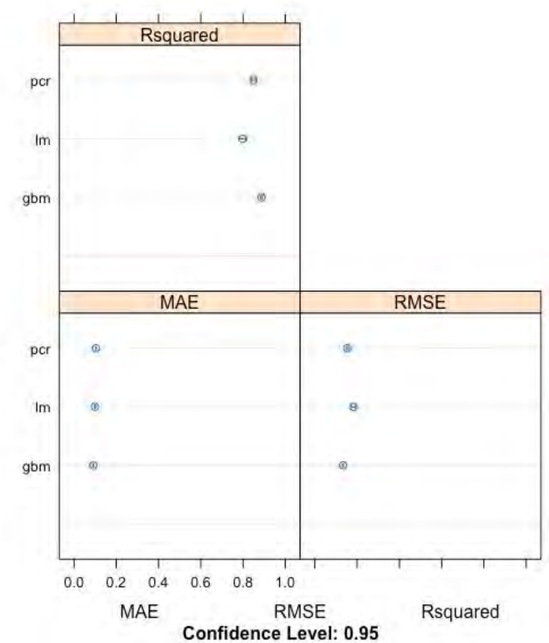
I chose a Gradient Boosting Model because it is well-suited for datasets of high multicollinearity, which makes sense for this dataset due to it having multiple variables describing the same feature. I compared it to a Principal Component Regression model, which is good for handling skewed data, as well as for handling the many related variables in the dataset.

Results

I used an easy multiple linear model, with predictors not log-transformed, as my basis for comparison. With the highest R^2 and the lowest RMSE, the Gradient Boosting model performed the best, with an initial RMSE of 0.13 and a post-accuracy test RMSE of 0.04.

Prediction Results (Numbers in parentheses are post-accuracy testing):

Linear	R^2= 0.7982	RMSE= 0.1835 (0.1966)
GBM	R^2= 0.8870	RMSE= 0.1331 (0.0429)
PCR	R^2= 0.8489	RMSE= 0.1534 (0.1365) *At 37 components



Discussion

When simply using bare-bones data that has minimal transformations, the multicollinearity-controlling capabilities of the GBM yield the best results.

That's not to say that the PCR model cannot be optimized. The PCR model works by making new predictor variables, or principal components, as linear combinations of the initial variables. If I were to do this project again, I would experiment with the optimal number of components in this PCR model.

I would also experiment with more treatments for outliers. I chose to not to delete outliers, instead choosing to attempt Log +1 transformations. When I log-transformed highly skewed variables, a few improved dramatically, some improved very little, while many seemingly became even less normally-distributed. This hurt the Linear model's performance most. In addition to treating outliers, I would also engineer more features to increase the performance of the linear model.

References

Boye, Paul & Mireku-Gyimah, Daniel. (2018). Principal Components Regression Model for Estimating the Price of a Housing Unit.

Winky K.O. Ho , Bo-Sin Tang & Siu Wai Wong (2020): Predicting Property Prices with Machine Learning Algorithms, Journal of Property Research, DOI: 10.1080/09599916.2020.1832558