Kaggle Competition Final Report

Ames, Iowa Housing Prices
IS 6489: Statistical and Predictive Analytics
Meag Tessmann | Dec 8, 2019

This report describes a model created for the Kaggle competition found at: https://bit.ly/2Q4kuJQ. The goal is to practice feature engineering and using advanced regression techniques to predict housing prices in Ames, Iowa based on a training set of 1460 observations with 20 continuous, 14 discrete, 23 ordinal, and 23 nominal variables. Performance is measured by lowest log RSME on a hold out set once predictions are submitted through the kaggle site. A training set is provided which I used in aid for comparing models.

Data Cleaning

I started off by filling in missing values in both the training and test sets. For variables which are physical features, such as whether a house has an Alley, I imputed a No Alley category, assuming they simply did not have the feature.

For descriptive fields of house features, such as the square footage of a basement, I entered a 0 when it appeared the feature did not exist. There were 81 observations, for example, which had the majority of the basement variables as NAs.

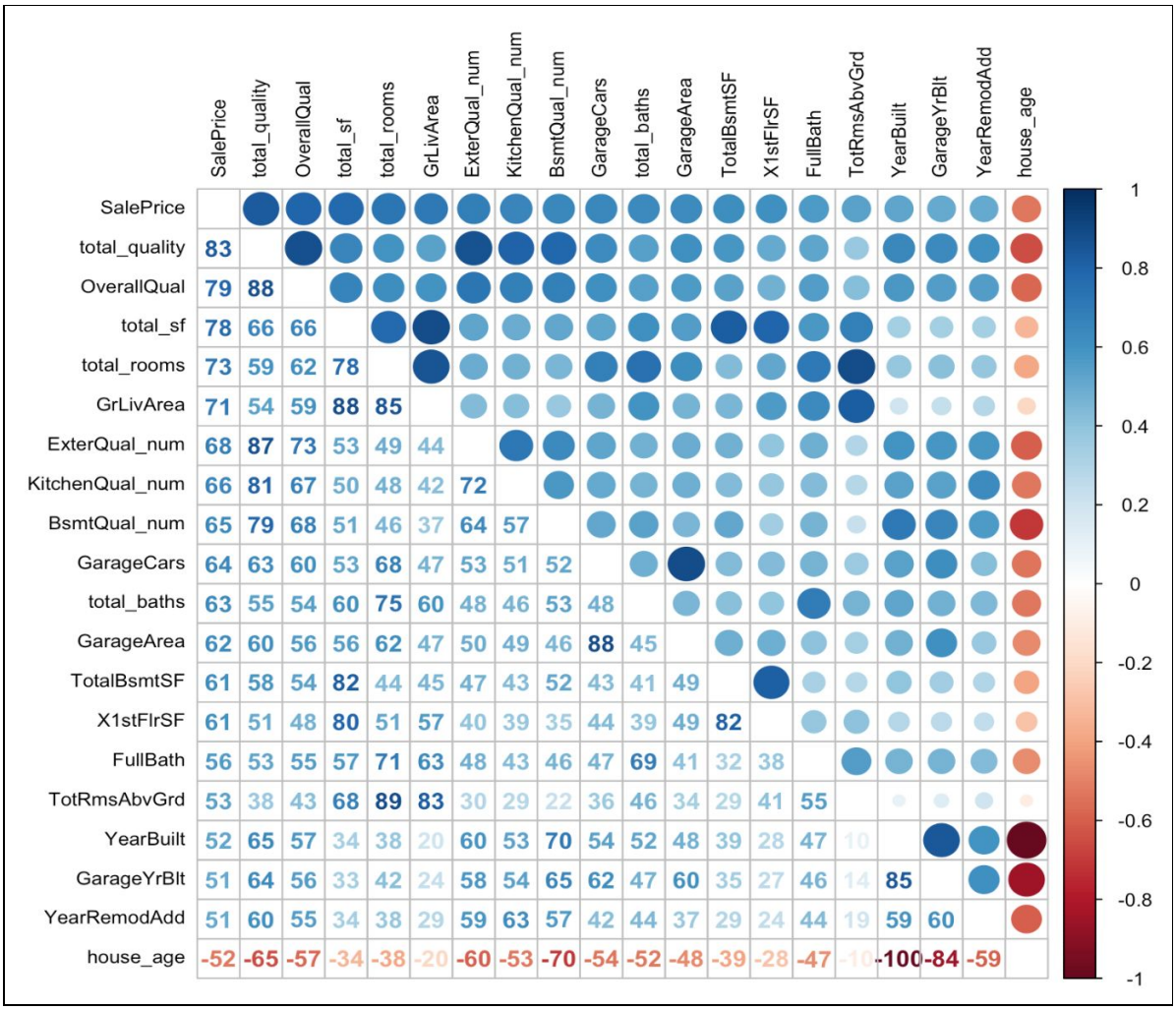
For other variables which every house should have, like the type of sale completed, I imputed the median for numeric variables or mode for categorical variables. There were a few variables I gave special treatment to - I imputed the year the house was built for the missing values which had a garage, but not a value for the year the garage was built.

There were a number of categorical variables which seemed to have an order, which I ordered as appropriate, such as “Gentle, Moderate, Severe” for the variable describing whether the land was sloped or flat.

Exploratory Data Analysis

A quick correlation plot of all of the numeric variables showed *Overall Quality* had a .79 correlation with SalePrice. The next more important variables were above grade living square footage and number of cars the garage had. I made a note of these for creating new variables. I saw some collinearity between a few variables - the square footage of the garage and the number of cars a garage could hold was a strong example. Another one is the year the house was built and the year the garage was built.

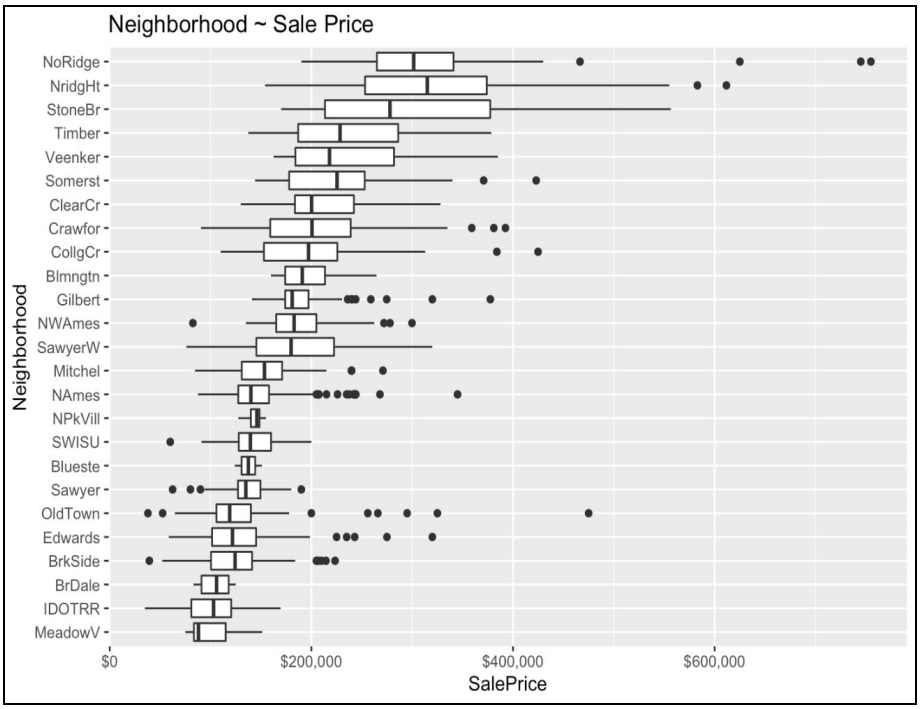
Correlation matrix of numeric variables with over .5 correlation with Sale Price in descending order. Correlation shown as percentage.



Additionally, I plotted all the numeric variables in bivariate plots against SalePrice to check for linearity. Six variables demonstrated a non-linear relationship. I discuss these below.

Of the categorical variables, neighborhoods had one of the highest variances when plotted against Sales Price. I use this later for interactions during model training.

Individual neighborhoods will be a good variable for interactions - spread of individual neighborhoods are small while variance between all neighborhoods is high.



Feature Engineering

I created 9 variables where I thought additional value could be gained by combining existing variables. I describe each of these below.

Total Quality

A numeric variable alternative to the overall quality variable provided. There are a series of quality variables, like fireplace quality or kitchen quality, which had an obvious order to them. For each of these variables, I created a new variable which I mapped to a numeric value given the following scale:

Quality Category	None/NA	Poor	Fair	Typical	Good	Excellent
Numeric Value	3	1	2	7	13	21

Notice poor and fair quality is weighted lower than if the house didn't have the feature at all. On the opposite side of the spectrum, Excellent has a mis-proportionally higher weighting. I then used these series of numeric variables to create a new 'Overall Quality' variable, which I made by multiplying the following together: exterior quality, kitchen quality, basement quality, garage quality, and overall quality. Having just a nice kitchen is not the same as also having a nice exterior, basement, and garage.

Total Squarefeet
A numeric variable accounting for total square footage, from both the basement and above ground levels.

Total Baths
A numeric variable summing of the bath variables: full baths and half baths from the above ground levels and the basement levels. Half baths are counted as .5.

Total Rooms
A numeric variable summing all of the room variables: total rooms above ground, total baths, and total cars in the garage.

Room to Bath ratio
A numeric variable dividing the number of bathrooms by the number of rooms.

House age
A numeric variable subtracting house's year built from year of sale.

Remodeled age
A numeric variable subtracting the house's last remodeled age from year of sale.

Yard
A numeric variable subtracting the square footage of the first floor from the total lot area.

Total porch area
A numeric variable summing the square footage of all the different types of porches: wood porches, open porches, enclosed porches, 3-season porches, and screened porches.

Statistical Model

Data Pre-processing

I removed 4 outliers which had a low sale price with a very high square footage, as square footage was one of the more strongly correlated variables.

I choose to use all of the variables available, since removing ones which appeared to have collinearity increased out of sample RMSE. I took the log of a few input variables which demonstrated a nonlinear relationship in a bivariate plot against the sale price: above grade square footage, the area of the lot, linear feet of street connected to the property, the area of masonry veneer, the garage area, the total square footage of the basement, and the square footage of wooden decks.

I one hot encoded all variables into a dummy matrix and removed near zero variance variables from the training set. In addition to all of the predictors mentioned, I one hot encoded three

interactions: total square footage by neighborhood, total rooms by neighborhood, and total quality by type of house.

Additionally, I centered and scaled all variables, as is required for the glmnet regression technique I choose.

Regression Model

In total, I trained on 1458 samples with 165 predictors. Comparing out of sample RMSE, a log model using glmnet regression performed the best at producing a model estimating Sale Price. The tuned model uses an alpha value of .383 and a lambda value of .008 when setting the seed to 123.

Model Performance

Measure of Success

In-sample RSME	Out-of-sample log RMSE	Kaggle Log RMSE	In-sample R^2	Out-of-sample R^2	Kaggle Rank
19655.68	0.1167199	0.12183	0.9399586	0.9148447	1237

The final model favored quality and square footage related variables.