Predicting house prices in Ames, Iowa

Fouad Yared
Johnny Mathis
Julio Pena

Hunter College | Stat 707 General Linear Models 2 | Spring 2020

Introduction

The price of a home is influenced by many different features. The following research looks at which features were most important in determining home prices in Ames, Iowa, a relatively small midwestern town (estimated population: 67,000) home to Iowa State University (U.S. Census Bureau). Sale prices were observed for January 2006 to July 2010, which overlaps with the Great Recession of 2008-2009 (which left millions without work and in difficult housing arrangements).

The features analyzed in this paper include a wide range of subjective and objective qualities regarding one's impression of the home's quality , the home's condition, the quality of the kitchen, the total area above ground (in square feet), the number of bedrooms, when the home was built, and when it was sold. Our goal was to use different machine learning models to find a model that best predicts housing prices in Ames, Iowa.

According to the 2018 American Community Survey (U.S. Census Bureau), the median household income in Ames, Iowa is $46,127 and the percentage of people living in the same home a year ago is 59.4%. Of those over 25 years old, over 96% graduated high school and over 62% received a Bachelor's degree.

Preparing the data

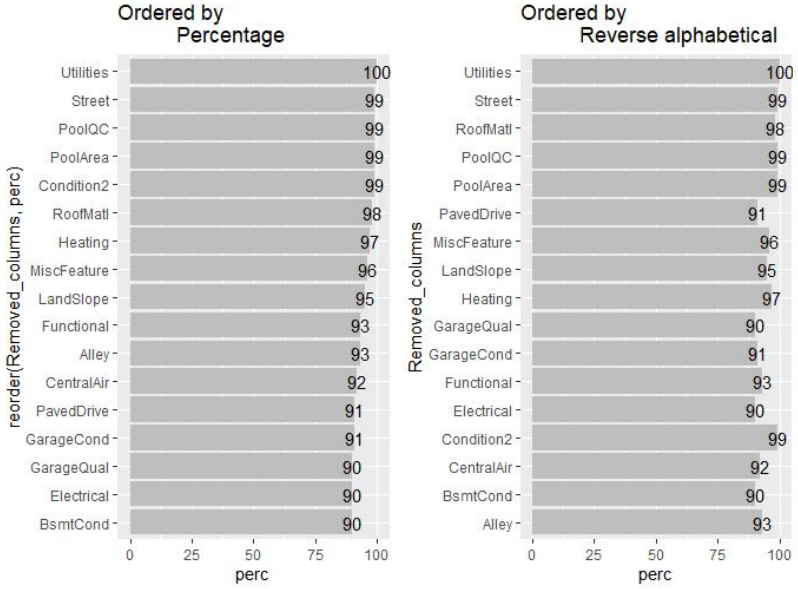
Variables were classified according to their type in R. Ratio features were treated as numeric, categorical features were treated as characters, and ordinal features were treated as ordered factors. Examples of numeric features are basement area (in square feet) and lot area. There was some ambiguity in how discrete and time variables should be treated. Specifically the number of bedrooms and the year something was built can be seen as ordinal or numeric. We decided to to treat them as ordered factors as the range of values was fairly limited for these features. Some levels of the factors were grouped together when there were too many levels and/or the counts in those levels was low.

Type	Classified as	Examples
Nominal	Characters	Building Type, Zoning Type, Type of Sale
Ordinal	Ordered Factors	Overall Quality (1 to 10), Number of bedrooms, Year Built
Ratio	Numeric	Basement Area, Lot Area, Sale Price

After classifying the variables, we looked at how common individual levels in categorical variables were. Since the train data set had 1,060 observations, we looked at removing variables where 90% (954) or more of the levels were the same. For instance, the categorical variable Street had two possible values: “Pavement” and “Gravel,” which appeared 1,054 and 6 times respectively. Since 1,054/1,060 of the observations were the same, we say that 99% were the same. The 17 variables that were mostly are shown in the bar plot below. We decided to retain five of these features as we wanted to test them in our models as they may have still been important. These are CentralAir, Electrical, Functional, LandSlope, and PavedDrive. Another feature, ScreenPorch, was removed because 92% of its values were the same, even though it was a numeric variable. There were a few other features that appeared multiple times that were removed because the second feature did not add

more information when compared to the first feature. Specifically, these are BsmtFinType2, BsmtFinSF2, and Exterior2nd. In total, 16 features were removed before modeling began.

Categorical Variables where 90%+ of cells have the same value



We next wanted to reduce the number of levels in the categorical variables. If a level in a categorical variable appeared less than 5% of the time (53 observations), then we would try to group it with another level, as shown in the set of tables below. The purpose of reducing levels was to limit the variation in predicting sale price from rare observations. In the top set of tables below, BldgType originally had five levels. These were conflated so similar features (2 family and duplex, the two Townhouse features) were added together. The bottom set of tables were conflated so the infrequently appearing irregular items (IR1, IR2, and IR3) became the “irregular” level. Overall 29 categorical variables had their levels reduced.

How levels in categorical variables were reduced (BldgType, LotShape)

Before	BldgType	1Fam	2fmCon	Duplex	Twnhs	TwnhsE
	Freq	879	24	44	34	79
After	BldgType	one_family	two_family	townhouse		
	Freq	879	68	113		

Before	LotShape	IR1	IR2	IR3	Reg	
	Freq	363	28	7	662	
After	LotShape	irregular	Regular			
	Freq	398	662			

The next step was to look at the variables with missing data. As noted in the Data Description, NA or null values in this data set usually meant the absence of a specific feature. For instance, the feature PoolQC had 1,055 values for NA and 5

non-NA. With this context, NA means the house doesn't not have a pool (both PoolQC and PoolArea were previously removed). Similarly, NA for basement or garage features meant there was no basement or garage. There were two features where NA did not mean a missing value. The values in these features, MasVnrArea and LotFrontage, were imputed with the median values of their Zoning Type and HouseStyle. Dealing with missing values allows us to retain observations when running various models.

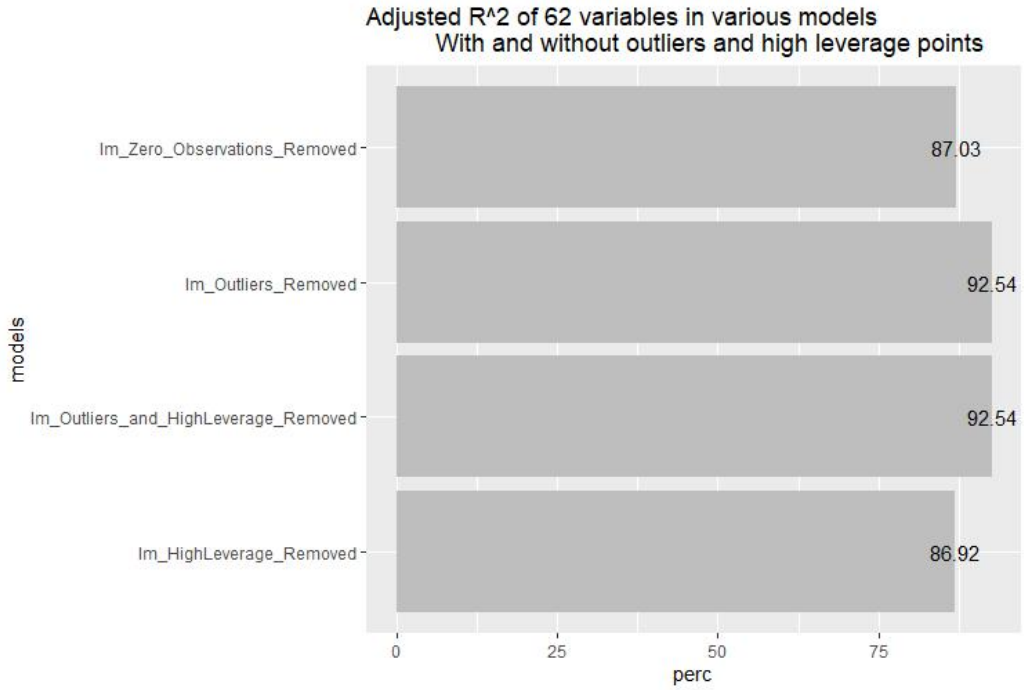
We were able to add features that had similar aspects. For instance, HalfBath and FullBath (above ground and in the basement) were four separate features. They were added to create one feature TotalBath. Also, we split a few of the yearly predictors into 10 year and ultimately 30 year buckets.

In order to improve our ability to explain the variation in sale price, we looked at univariate and multivariate outliers along with high leverage points. To find univariate outliers and gather Z-scores, we centered and scaled every numeric feature. We looked at the Z-scores for four features: LotArea, LotFrontage, GrLivArea, and SalePrice. These features had five, one, four, and four Z-scores above 5 (a conservative threshold), respectively. Since there was a lot of overlap, ten observations were ultimately selected (for instance, observation with Id 1299 was an outlier for the first three features listed). It is likely these observations were not data errors, but they were rare when compared to the values of other observations. We decided to remove these ten observations to improve model performance.

To find multivariate outliers, we looked to see whether observations were close to the multivariate center of the distribution using the Mahalnobis D for the numeric columns. Z-scores were subsequently gathered for these values. Only one observation had a Z-score above 5.0 and it was previously removed (id 1299).

R provides many ways of finding potentially high leverage points. We focused on two standardized values: Cook's distance and hat values. Cook's distance measures the affect of one observation on all fitted values (Pennslvania State University 2018, Boston University School of Public Health 2016). Hat values show the leverage of each observation based on the mean values of the predictors (Boston University School of Public Health 2016, Rodríguez 2020). Observations were flagged when Cook's distance was greater than (4/n) or when hat values were greater than 2*(p+1)/n where n is the number of observations and p is the number of parameters. We identified 50 high leverage points: 35 observations had a Cook's distance above the threshold, 23 had hat values above the threshold, and 8 had both.

To compare the output of models with and without outliers and high leverage points, four multiple regression models were run with 62 variables and are compared in the below chart. Before removing outliers and high leverage points, our adjusted R^2 value was 87.03%. When the 10 outliers were removed, it improved to 92.54%. When both outliers and high leverage points were removed, the adjusted R^2 value remained at 92.54%. When only high leverage observations were removed (and outliers were retained), the model actually performed worse than retaining the initial model where all observations were retained.

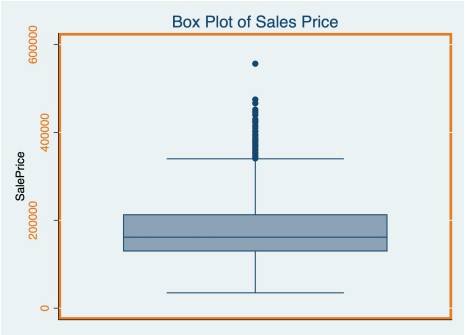


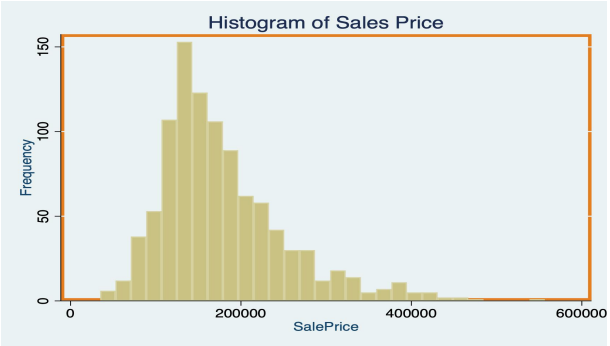
Using different models to conduct variable selection

Before fitting different models, we looked at the distribution of SalePrice. The below histogram and box plot show that the distribution is skewed to the right. Instead of regressing our features on the log of our response, we decided to regress on SalePrice without a transformation. This would allow us to more easily interpret the coefficients.

Dependent Variable - Sales Price

Min = 34,900 / Median = 161,625 / Mean = 178,229 / Max = 556,581





Forward Stepwise Regression

We started the variable selection process with Forward stepwise regression, a commonly used procedure when the number of features (independent variables) is more than 40 due to its computational efficiency. Starting with an empty model, predictors are added one at a time depending on the criteria chosen. A variable is added to the model if it is minimizes or maximizes a chosen criteria and exceeds a cost function. Forward selection is a greedy process because it looks for the next best feature one at a time, as opposed to finding the optimal set of features.

The criteria chosen for our forward selection model is the Akaike Information Criteria (AIC). AIC is "defined when models can be fit with maximum likelihood" (James, 2013). Using AIC in the forward selection process allows us to find the number of variables in a model that minimize the AIC score. The following formula was used to find the lowest AIC:

$$AICp = n \ln SSEp - n \ln n + 2p$$

Thirty-two features produced the lowest possible AIC value of 20897.02. These variables were organized into the following three categories: inside the property, outside the property, and about the property.

Inside the Property (12)	**Outside the Property (10)**	**About the Property (10)**
BsmtFinSF1	Exterior1st	Condition1
BsmtExposure	ExterCond	Functional
BsmtQual	GarageCars	GarageYrBlt_decade
BsmtFinType1	LandContour	MSSubClass
BsmtUnfSF	LotArea	Neighborhood
CentralAir	LotConfig	OverallQual
FireplaceQu	MasVnrArea	OverallCond
GrLivArea	OpenPorchSF	SaleType
HeatingQC	WoodDeckSF	SaleCondition
KitchenAbvGr	X3SsnPorch	YearBuilt_decade
KitchenQual		
TotalBsmtSF		

A multiple linear regression (MLR) model was explored with SalePrice as the dependent and the aforementioned 32 predictor variables as the independent variables. The model produced an adjusted R^2 of .9265 and an MSE of 4.0013e+08. Six predictors had a p-value greater than 0.05. A p-value below 0.05 likely rejects our null hypothesis that the coefficient for that predictor is zero. Conversely, a p-value above 0.05 means a coefficient for that predictor may not be different than zero.

These predictors were FireplaceQu, LandContour, BsmtUnfSF, ExterCond, MasVnfArea, and OpenPorchSF. Their p-values can be found in the below table.

	Pr(>|t|)
FireplaceQuaverage_or_worse	0.235654
FireplaceQuno_fireplace	0.419409
LandContournot_normal	0.052774
BsmtUnfSF	0.052281
ExterCondaverage_or_below	0.07875
MasVnrArea	0.074576
OpenPorchSF	0.095618

First Attempt eliminating predictor variables after analyzing p-values

We were then interested in seeing how the adjusted R^2 and MSE would be affected if the six predictors were removed from the model.

The MLR model with the 26 remaining predictor variables had an adjusted R^2 of .9254 and an MSE of 4.0583e+08. After fitting this model, we noticed three predictors that had p-values greater than 0. 05. Those variables were BsmtFinType1, X3SsnPorch, and LotConfig.

Second attempt eliminating predictor variables after analyzing p-values

With the adjusted R^2 and the MSE barely affected, the MLR was re-examined with the 23 remaining predictor variables. Now, our MLR model has an adjusted R^2 of .9243 and an MSE of 4.1171e+08. Another three predictors had p-value s greater than 0.05. These predictor variables are: MsSubClass, CentralAir, and GarageYrBlt.

Last attempt eliminating predictor variables analyzing P-values

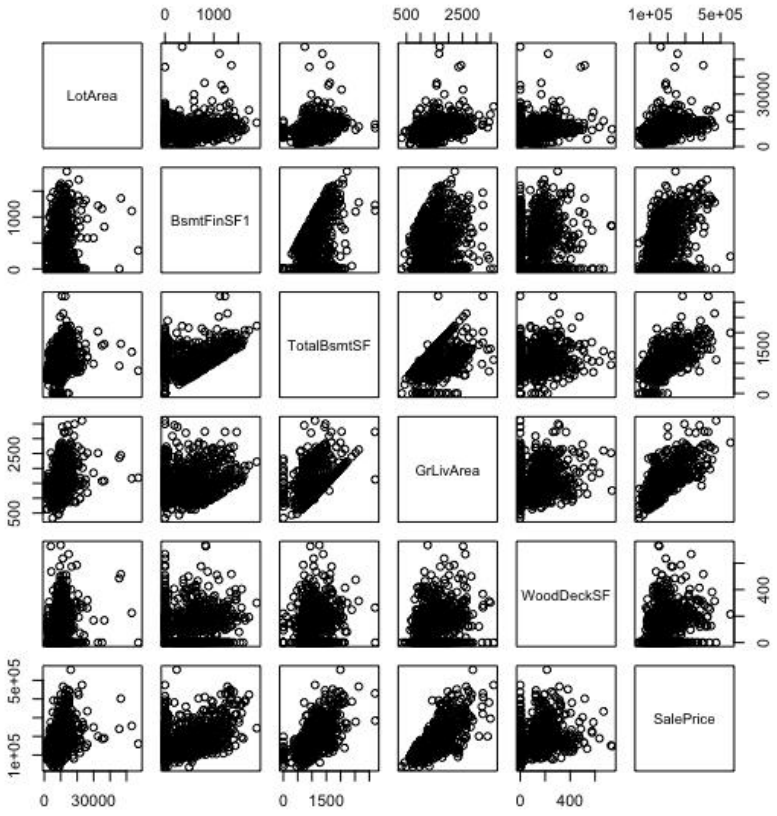
Analyzing our MLR Model with the 20 remaining predictor variables, the adjusted R^2 was .9235 and the MSE was 4.1647e+08. All of the predictor variables have at least one parameter in this updated model that were significant with an alpha level of .05.

Variable selection: Using Forward Selection and MLR		
	Adjusted R^2	MSE
First attempt	0.9265	4.00E+08
Second attempt	0.9254	4.06E+08
Third attempt	0.9243	4.12E+08
Fourth attempt	0.9235	4.16E+08

Checking Collinearity and Multicollinearity

Collinearity among numeric features was evaluated with both a correlation matrix and a scatterplot matrix. Multicollinearity was assessed with the Generalized Variance Inflation Factor (GVIF), where terms were squared to get the vif. A conservative threshold for assessing multicollinearity is a vif score of 3 or greater.

	LotArea	BsmtFinSF1	TotalBsmtSF	GrLivArea	WoodDeckSF	SalePrice
LotArea	100%	18.6%	29.6%	36.4%	17.9%	37.0%
BsmtFinSF1	18.6%	100%	42.0%	9.7%	16.5%	35.5%
TotalBsmtSF	29.6%	42.0%	100%	38.9%	23.6%	62.1%
GrLivArea	36.4%	9.7%	38.9%	100%	26.5%	73.6%
WoodDeckSF	17.9%	16.5%	23.6%	26.5%	100%	33.7%
SalePrice	37.0%	35.5%	62.1%	73.6%	33.7%	100%



Referencing the correlation and scatterplot matrices, there are 2 pairs of predictor variables that are fairly correlated. The features BsmtFinSF1 and TotalBsmtSF are correlated with a correlation coefficient ρ of 0.42. Similarly, GrLivArea and LotArea are correlated with a ρ of .364. TotalBsmtSF and GrLivArea are both strongly correlated with SalePrice. as they have a ρ of 0.621 and 0.736 respectively.

Based on the generalized vif value above 3, SaleType is highly correlated with other predictors. Due to evidence of possible multicollinearity, a model without the following variables will be considered : BsmtFinSF1, LotArea and SaleType.

Further analysis of this MLR model with 17 predictor variables

After removing 3 variables due to possible multicollinearity, our adjusted R^2 decreased to 0.91 and the MSE increased to 4.9092e+08. HeatingQC is the only variable that is not significant with an alpha level of .05. As a result, HeatingQC was removed from the model.

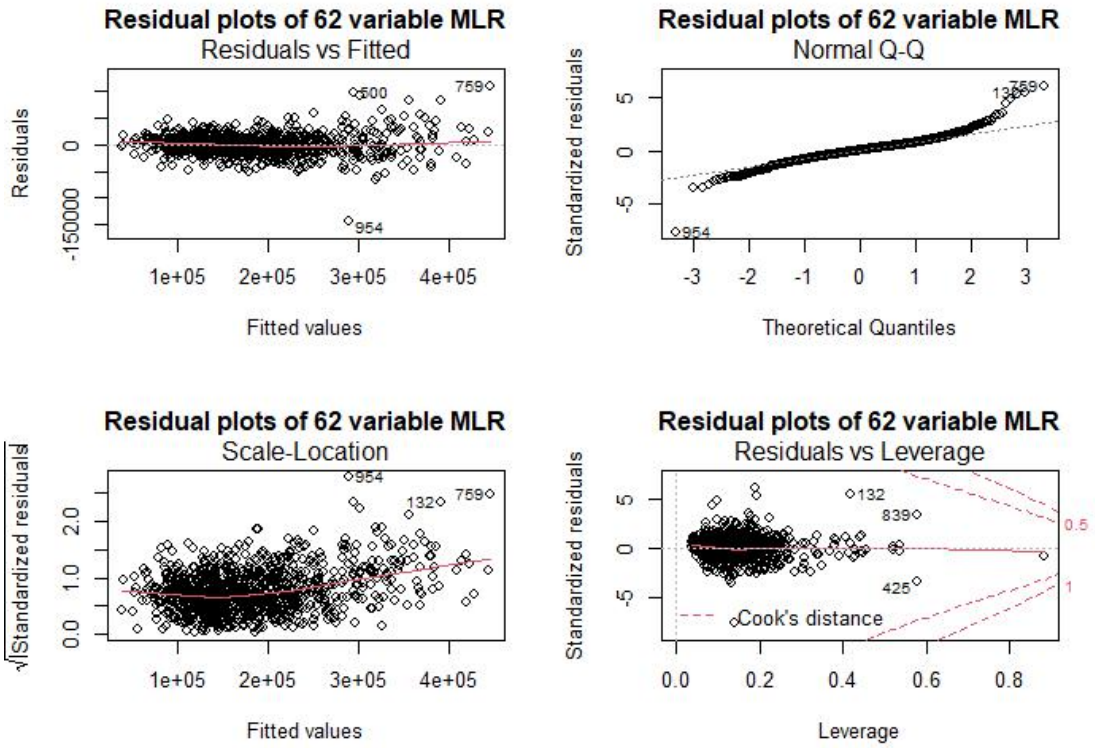
The final model, using this method, has an adjusted R^2 of 0.91 and MSE of 4.9265e+08. This MLR model will contain the following 16 predictor variables:

Inside the Property (6)	**Outside the Property (3)**	**About the Property (7)**
BsmtExposure	Exterior1st	Condition1
BsmtQual	GarageCars	Functional
GrLivArea	WoodDeckSF	Neighborhood
KitchenAbvGr		OverallQual
KitchenQual		OverallCond
TotalBsmtSF		SaleCondition
		YearBuilt_decade

Multiple regression model

We also used multiple regression models outside of forward selection to gain a general understanding of which features had p-values below 0.05. Multiple linear regression, one of the more traditional and interpretable models, assumes a linear relationship between the predictors and the response. Another assumption is to have constant error variance (homoscedasticity) and a lack of correlation among the error terms. In a previous section, we dealt with univariate and multivariate outliers and high leverage points. Collinearity and multicollinearity will be assessed in the final model.

When regressing all predictors against SalePrice in the 62 variable model, the adjusted R^2 is 0.9199 and the MSE is 406,084,504. There are several variables that have p-values below 0.05. We further reduced these variables by only including those that are generally known to affect housing prices. The lists consists of LotArea, Neigborhood, Condition1 (proximity to a positive or negative external feature), OverallQual, OverallCond, Exterior1st, BsmntQual, GarageCars, TotalBsmtSF, X1stFlrSF (area of 1st floor), X2ndFlrSF (area of 2nd floor), LowQualFinSF, KitchenAbvGr, KitchenQual, and Functional (whether deductions regarding potential renovations are necessary). Some of the features identified seem universally important (e.g., Neighborhood is generally important) while some features may be specifically important to Ames, Iowa (e.g., GarageCars is important as living in Ames, Iowa may necessitate having a car while it is not necessary in NYC).



The relationship between the residuals and the fitted values seems to be somewhat linear. In the Normal Q-Q plot, the relationship between the theoretical quantities and standardized residuals looks normal. The Scale-Location plot shows whether there is constant variance among residuals. Since the red line is not flat, we do not seem to have constant variance among residuals. (This will be tested in the final model.) Additionally there are a few observations we may want to further inspect due to their high leverage points using Cook's distance. As a reminder, we identified 50 high leverage points using both Cook's distance and the hat values, but chose not to remove them.

Although the adjusted R^2 value and MSE are strong, we wanted to reduce the set of predictors to make the model more interpretable.

Lasso and Ridge Regression

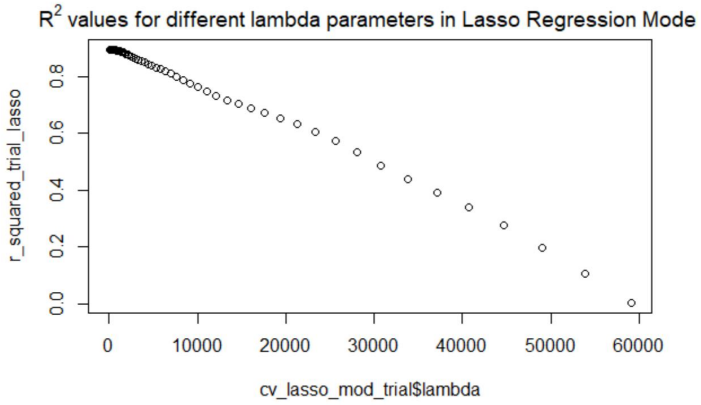
Both lasso regression and ridge regression are very similar to multiple linear regression. They rely on the least squares estimates and add a shrinkage penalty λ. The shrinkage penalty reduces the B terms to exactly zero (with lasso) or close to 0 (with ridge). While lasso regression helps with feature selection, ridge regression helps with multicollinearity. As λ increases, our coefficients further shrink towards zero. When λ=0, the least squares estimates are produced. When λ=∞, all B=0. (There is a lot of variation in values for λ, depending on the data set. For instance, values may be in the hundreds or thousands.) We use cross-validation to find the right level of λ.

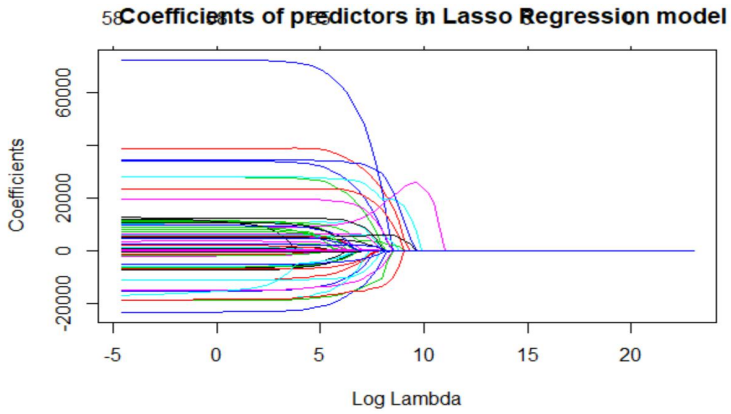
For lasso regression on the 62 variable model, we split the train set using most variables and ran a cross validation technique that standardized the variables. Our cross validation produces a λ of 294.1 as this is the lambda value associated with the smallest MSE. We then fit the model and predicted values. Since lasso regression brings some variables or levels of variables to exactly zero, we chose to include all variables that had at least one non-zero coefficient for a specific level. The R^2 value is 90.76% and the MSE is 516,962,202 for the 62 variable Lasso Regression model.

Lasso Regression Equation

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

Then we performed a “Relaxed” lasso regression. Since lasso brings all variables towards zero, we re-ran lasso with only the variables that have non-zero coefficients to get Bs for interpretation. The reduced lasso regression left us with 53 predictors. The below chart shows the R^2 values for the lasso regression model as λ. Additionally, as λ increases, the coefficients shrink to exactly zero. The R^2 value is 90.96% and the MSE is 517,879,541 for the 52 variable Ridge Regression model (where at least one level of each variable has a non-zero coefficient).



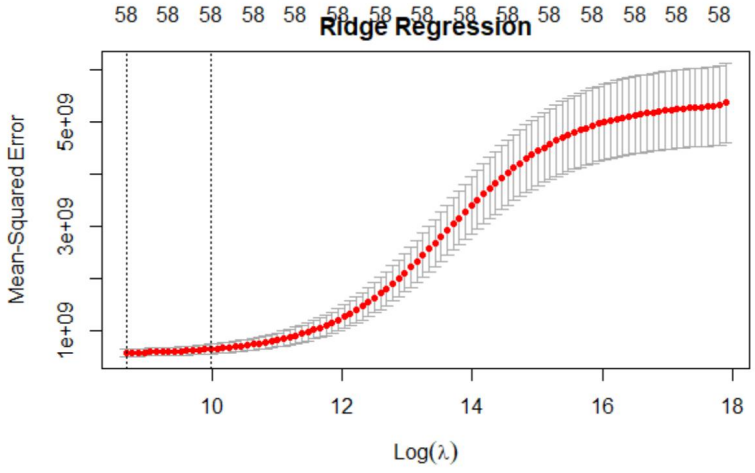


Ridge Regression on the 62 variable model performs similarly. We used a cross validation technique and standardized the values. Cross validation produced a λ of 5909.5 (the λ for a non-standardized model is much greater). We then fit the model and predicted values. The R^2 value is 90.37% and the MSE is 599,633,800 for the 62 variable Ridge Regression model.

Ridge Regression Equation

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

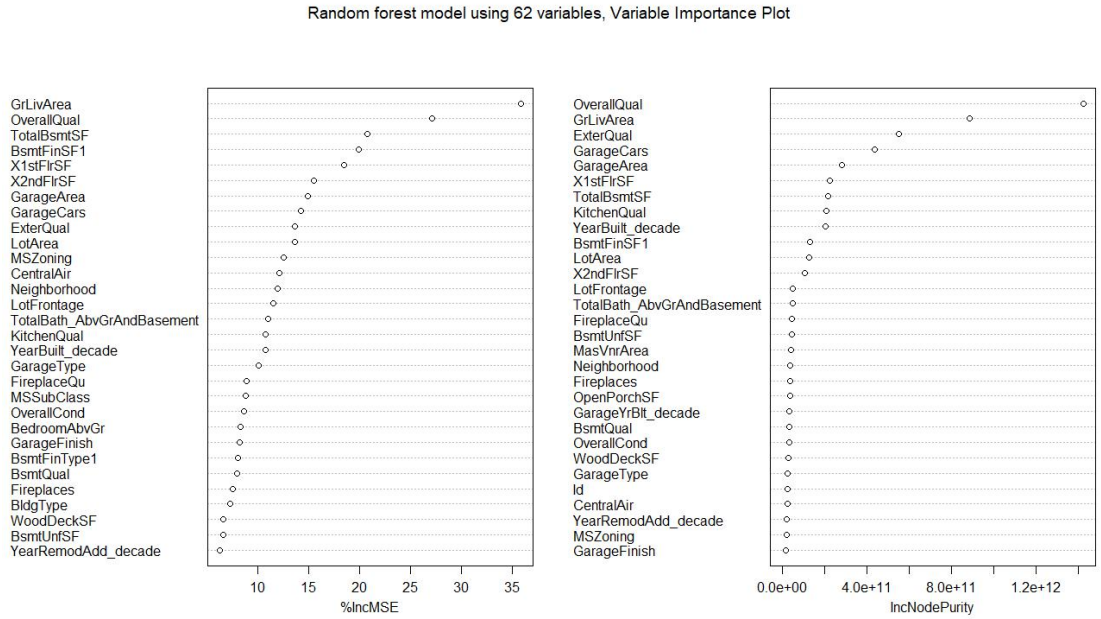
In the chart below, the lowest point in the curve indicates the optimal value of lambda as the log value of lambda minimizes the cross-validation error.



Random Forest

Random forest models are constructed with decision trees which split various features on some value (e.g., observations may be split on GrLivArea where values greater than or equal to 1500 go on one side and values less than 1500 go on the other side) and further splits them on other values. The SalePrice in the terminal node is then averaged. Random forest works by selecting a random predictor for the next level (usually in the set of √p) and running many trees to determine which features are most important in predicting home prices. This also decorrelates the features. It's worth noting that random forest models do not have an explicit shape

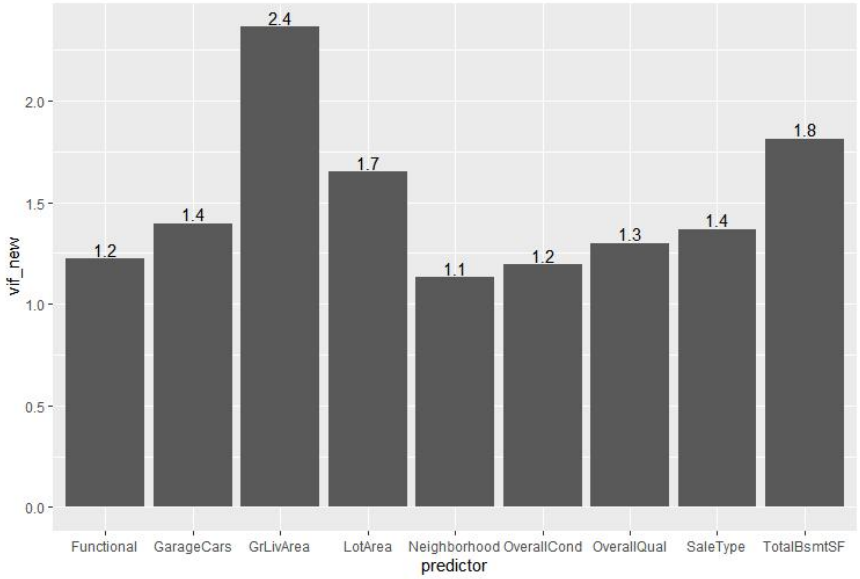
like linear models; they are non-parametric. The R^2 value is 88.58% and the MSE is 620,886,793 for the 62 variable Random Forest model.



Deciding on the most important predictors

Relying on different models each of us ran, the group agreed on nine variables that accounted for an R^2 of 88.8%. The benefit of using different models is that it allowed to identify variables that were not significant in certain models yet they still performed very well (e.g., GrLivArea performed very well in Random Forest but not in Lasso Regression). There was an additional six variables a subset of us were interested in. The nine variable model consisted of GarageCars, Neighborhood, OverallCond, OverallQual, Functional, LotArea, SaleType, TotalBsmtSF, and GrLivArea. The fifteen variable model had an R^2 of 89.8%. (The additional six variables were BldgType, KitchenQual, KitchenAbvGrd, SaleCondition, and Condition1.)
Although all the variables were significant in the fifteen variable model, we decided to go with the parsimonious one.

We then evaluated multicollinearity for the nine variable model. We squared the generalized vif. All vif values were below 3 (see chart below) and were retained.



Model Validation

The purpose of model validation was to see how good our various models perform on new (test) data. Cross-validation methods were used for multiple linear regression, lasso regression, and ridge regression. Using cross-validation on the rpart() Random Forest model led to a very high MSE, so the randomforest() function was used without cross-validation (although 500 trees were created). The table lists the evaluation criteria when comparing the different models.

Evaluation criterion	Formula
Bias	Mean(Yhat - Y)
Maximum Deviation	Max(abs(Yhat - Y))
Mean Absolute Deviation	Mean(abs(Yhat - Y))
Mean Squared Error	Mean(Yhat - Y)^2

Evaluating the models

Generally, multiple linear regression, lasso and ridge regression performed similarly when using cross validation techniques (see rotated table on next page). Random forest performed the worst on all measures, but it helped us identify GrLivArea as very important in reducing MSE. Although it does not have the lowest MSE, we chose Multiple linear regression as our model of choice because of its ease in interpreting and high R^2. (A relaxed lasso regression was not necessary as none of the nine predictors were set to zero in the original lasso model.)

9 variable model	Bias (using Mean)	Max Deviation	Mean Absolute Deviation
Multiple linear regression	-151	115,611	18,263
Lasso regression	762	150,183	17,383
Ridge regression	697	150,998	17,350
Random forest	-11,298	268,938	23,579

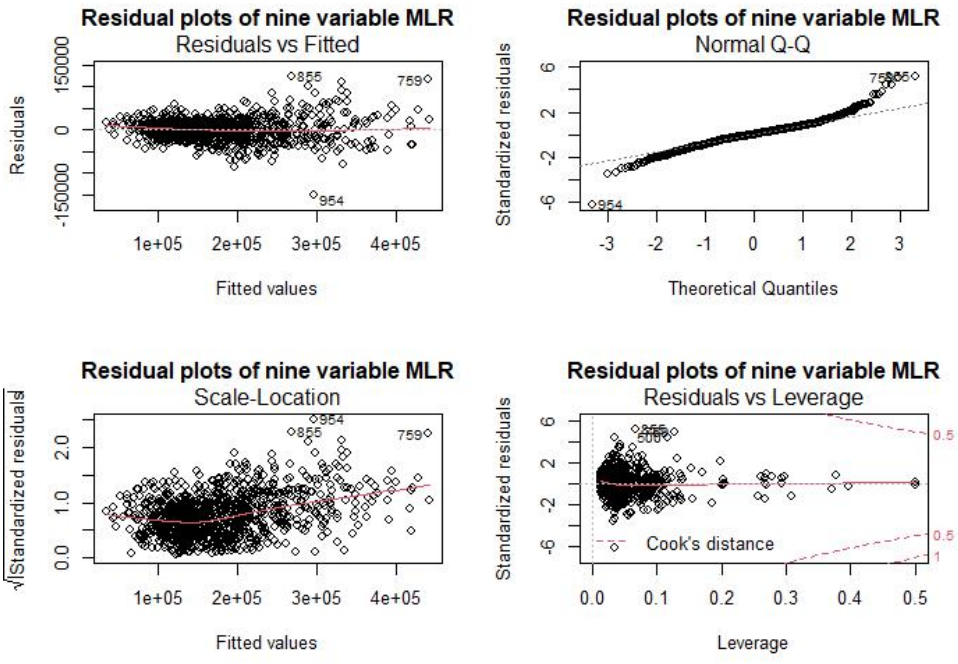
9 variable model	MSE	R^2
Multiple linear regression	615,889,941	88.8%
Lasso regression	602,170,342	88.2%
Ridge regression	607,638,193	88.0%
Random forest	765,709,896	85.9%

Interpreting the Final Multiple Linear Regression output:

All nine variables in the final model were significant with p-values below 0.001, as shown in the table below.

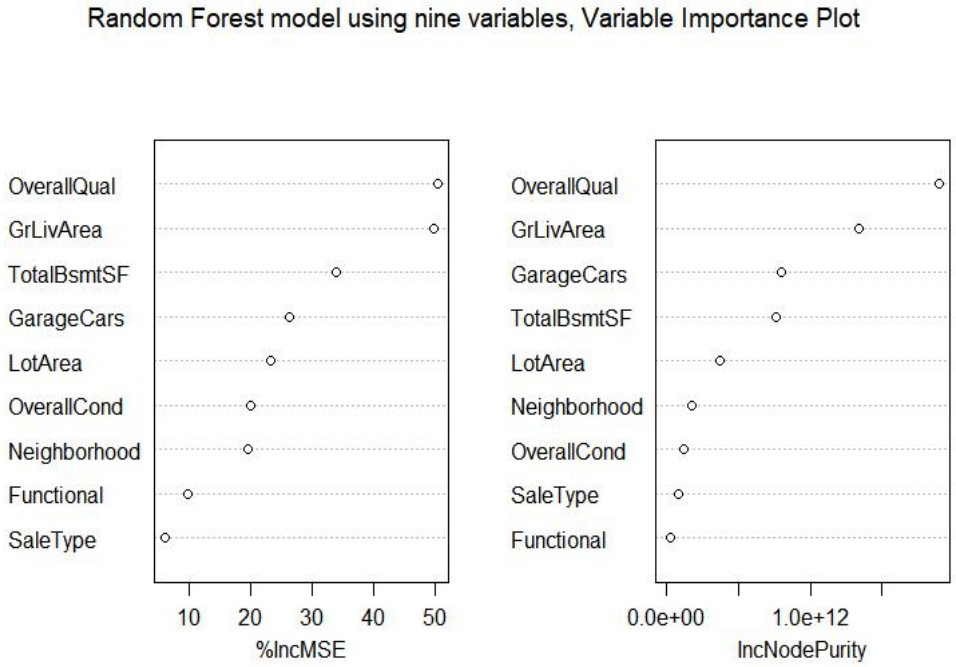
Feature	Df	SumSq	MeanSq	F value	P-value	Sig level
GarageCars	4	2.82E+12	7.04E+11	1150.9473	< 2.20E-16	***
GrLivArea	1	1.12E+12	1.12E+12	1831.4626	< 2.20E-16	***
OverallCond	7	1.65E+11	2.35E+10	38.4422	< 2.20E-16	***
OverallQual	9	6.07E+11	6.74E+10	110.1732	< 2.20E-16	***
LotArea	1	7.14E+10	7.14E+10	116.7603	< 2.20E-16	***
SaleType	1	5.64E+10	5.64E+10	92.2494	< 2.20E-16	***
Functional	1	1.87E+10	1.87E+10	30.5669	4.116E-08	***
TotalBsmtSF	1	1.12E+11	1.12E+11	183.4115	< 2.20E-16	***
Neighborhood	24	1.29E+11	5.38E+09	8.7926	< 2.20E-16	***
Residuals	1000	6.12E+11	6.12E+08			

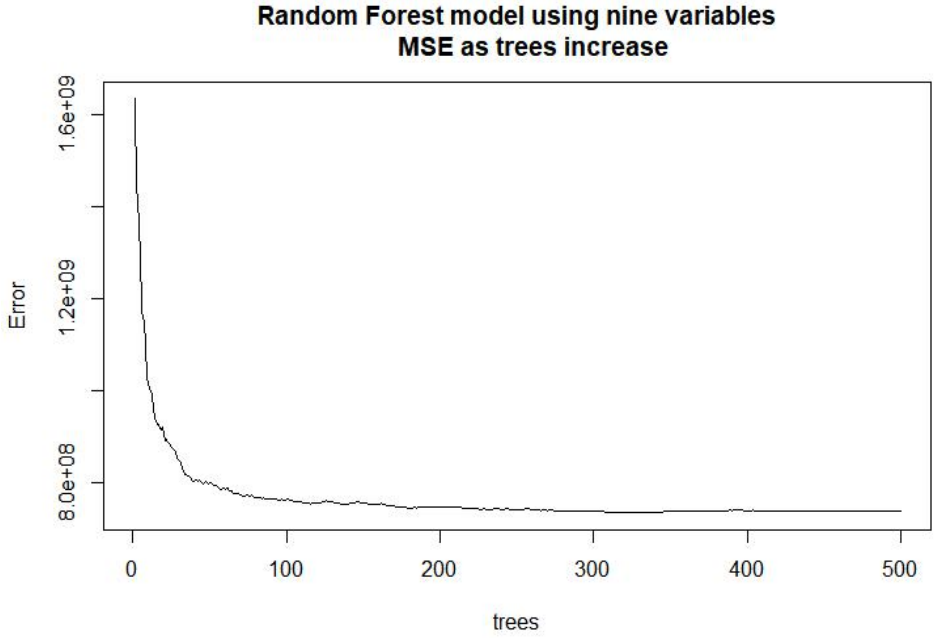
The residual plots show that the residuals seem fairly linear. The qq-plot shows the standardized residuals deviate from normality at the two ends of theoretical quantiles. Since the red line in the Scale-Location plot, there may not be constant variance among the residuals. We tested this with the Breush-Pagan test (using the 9 variable model). The null hypothesis is that the model has constant variance among residuals. Since the result was less than 0.05, we reject the null hypothesis and assume heteroscedasticity. The final plot shows leverage points using Cook's distance.



Interpreting the Final Random Forest output:

The final random forest output shows us that most of the variables in the reduced model contribute greatly to MSE. The two most important variables in the model were Overall Quality (OverallQual) and Above Ground Living Area (GrLivArea). The second plot shows that MSE values converge around 8,000,000 as the number of trees increases.





Conclusion

The objective of this research was to find the features and models that best predict housing prices in Ames, IA. The most important features were related to square footage, the quality and condition of the home, its neighborhood, the number of cars in the garage, and whether the house is functional. Instead of focusing on accuracy and improving our R^2 (88.8%), our nine variable final model is easily interpretable.

Future research can look into meeting the different assumptions of linear regression and seeing whether other models are a better fit for this type of problem, comparing the results of this analysis for Ames, Iowa similar analyses in other parts of the country (e.g., more urban and/or non-midwestern settings).

References

Boston University School of Public Health. (2016). Regression Diagnostics. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression7.html

James, G., Witten, Daniela, W., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning: with applications in R. New York: Springer http://faculty.marshall.usc.edu/gareth-james/ISL/

Pennslvania State University, The. (2018). Identifying Influential Data Points. https://online.stat.psu.edu/stat462/node/173/

Rodríguez, G. (2020). Regression Diagnostics https://data.princeton.edu/wws509/r/c2s9

U.S. Census Bureau Quickfacts: Ames City, Iowa (2018). American Community Survey, 2018 One year estimates). https://www.census.gov/quickfacts/amescityiowa