Predicting House Prices in Ames, Iowa

Fouad Yared

Johnny Mathis

Julio Pena

Introduction

Our objective and methodology

Find the features that have the greatest influence on home prices, make predictions, and evaluate which model(s) perform the best.

1. Identify data types
2. Clean predictors
3. Use different models to conduct variable selection
4. Build the final model and evaluate

Part I: Data Preparation

How variables were classified

80 total variables

23 are nominal, 23 are ordinal, 14 are discrete, and 20 are continuous

Nominal variables: unordered categories | Ex: Neighborhood, SaleType

Ordinal variables: have order, unequal differences between levels | Ex: OverallQual

Discrete variables: obtained by counting | Ex: # of bathrooms

Numeric variables: obtained by measuring| Ex: Basement Square Feet

Part II: Cleaning Predictor Variables

Part II: Cleaning Predictor Variables

- Determine which categorical variables have little variation
  - Variables with one predominant level (90-95%+):
    - Utilities (100%), Street (99%), Condition2 (99%), PoolQC (99%), RoofMatl (98%)
    - 13 variables removed
- Reduce rare levels in categorical variables
  - Using research and intuition
    - LotShape regular / irr1 / irr2 / irr3 became regular / irregular
- Variables with missing data
  - Impute data with zoning & house style. Ex: LotFrontage. Mark some as no garage/no bsmt
- Combine variables, create new ones
    - Similar aspects: combined. TotalBath. Decades for Yearly data

Dealing with outliers, high leverage points

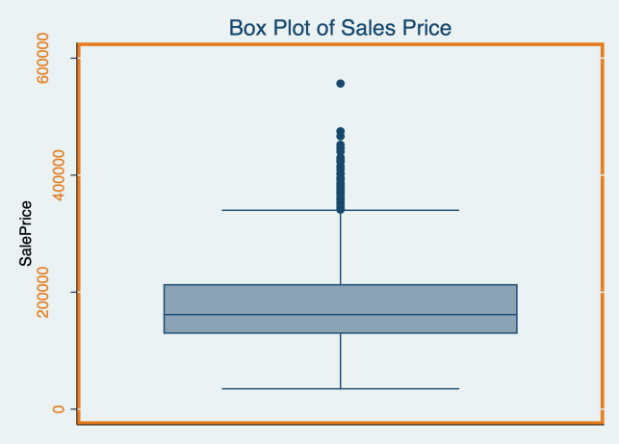
- Univariate outliers
  - Each numeric feature was centered and scaled
  - Z scores were analyzed to determine whether specific outliers were very different
  - Outliers observed were likely not a data error. Some observations had very large values
  - We chose to remove 10 univariate outliers with Z scores above/below 5
- Multivariate outliers
  - Mahalanobis D used: are observations close to the multivariate center of a distribution?
  - One variable was highlighted and it was already flagged with univariate detection
- High leverage points
  - Using cook's distance and the hat matrix, we identified 50 *potential* high leverage points
  - Observations flagged if Cook's distance > 4 / n or hat matrix > 2*(p+1)/n
  - Two data sets: 1) outliers removed and 2) both outliers and leverage points removed

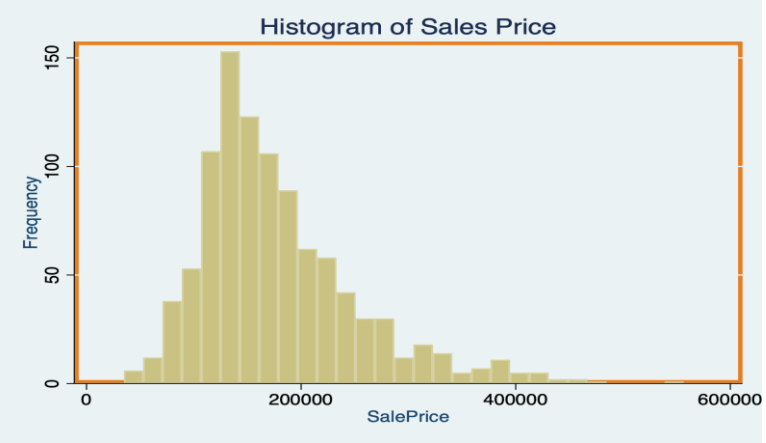
Part III: Model Selection

Exploratory data analysis

- Box plots, histograms, and scatterplots were created to analyze univariate and bivariate statistics

- Dependent Variable - Sales Price
  - Min = 34,900 / Median = 161,625 / Mean = 178,229 / Max = 556,581





Part III: Model Selection

- Forward Stepwise Selection
- Multiple Linear Regression Model
- Lasso Regression
- Ridge Regression
- Random Forest Regression

Forward selection: All models

- First Attempt
  - Summary of Forward Stepwise Regression
    - Adjusted R^2 : .9265
    - MSE: 4.0013e+08

- Second Attempt
  - This model excluded FireplaceQu, LandContour, BsmtUnfSF, ExterCond, MasVnfArea, OpenPorchSF.
    - Adjusted R^2 : .9254
    - MSE: 4.0583e+08

- Third Attempt
  - This model excluded FireplaceQu, LandContour, BsmtUnfSF, ExterCond, MasVnfArea, OpenPorchSF, BsmtFinType, X3SsnPorch, LotConfig.
    - Adjusted R^2 : .9243
    - MSE: 4.1171e+08
- Fourth Attempt
  - This model excluded FireplaceQu, LandContour, BsmtUnfSF, ExterCond, MasVnfArea, OpenPorchSF, BsmtFinType, X3SsnPorch, LotConfig, MsSubClass, CentralAir, GarageYrBlt.
    - Adjusted R^2 : .9235
    - MSE: 4.1647e+08

Remaining Predictor Variables with at least one Significant Parameter

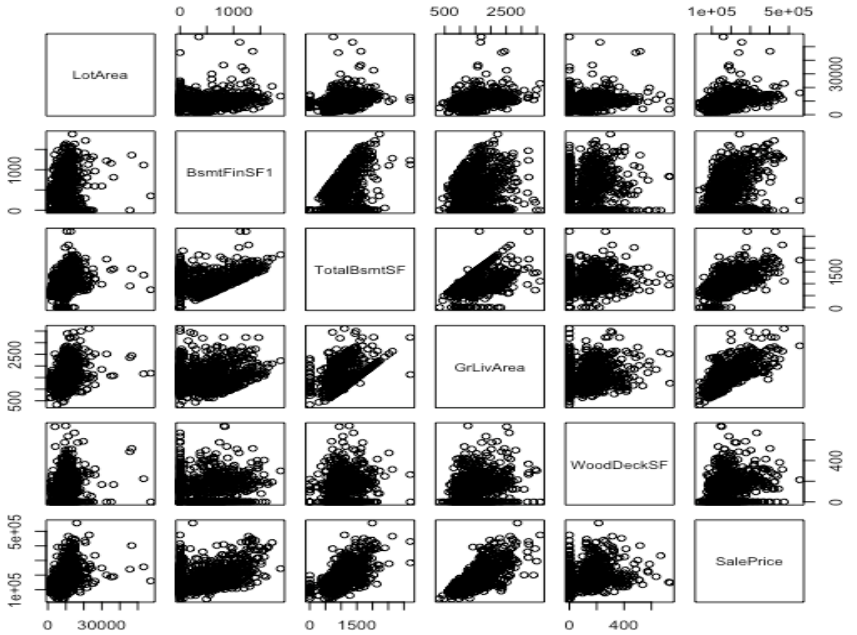
OverallQual
GrLivArea
Neighborhood
BsmtFinSF
SaleCondition
OverallCond
BsmtExposure
YearBuilt_decade
GarageCars
Functional

TotalBsmtSF
LotArea
SaleType
KitchenQual
Exterior1st,
BsmtQual
KitchenAbvGr
WoodDeckSF
Condition1
HeatingQC

Checking for Collinearity

	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%

	LotArea	BsmtFinSF1	TotalBsmtSF	GrLivArea	WoodDeckSF
Vif	1.21	1.24	1.46	1.34	1.11



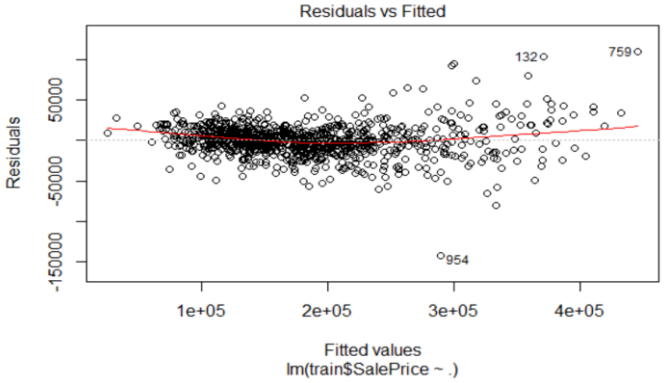
Multiple Regression Model

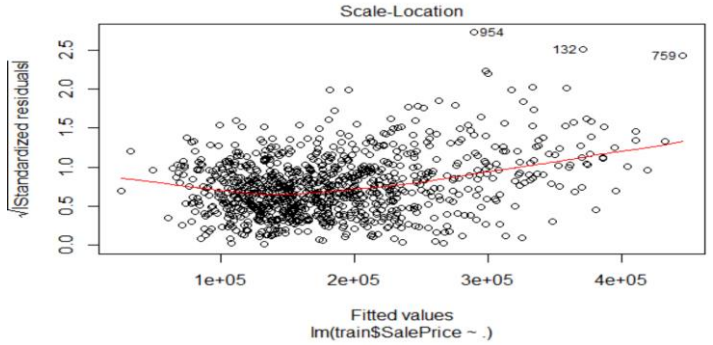
Multiple Linear Regression is useful for modeling the relationship between a numeric outcome or dependent variable (Y = Sales Price) and multiple explanatory or independent variables (X).

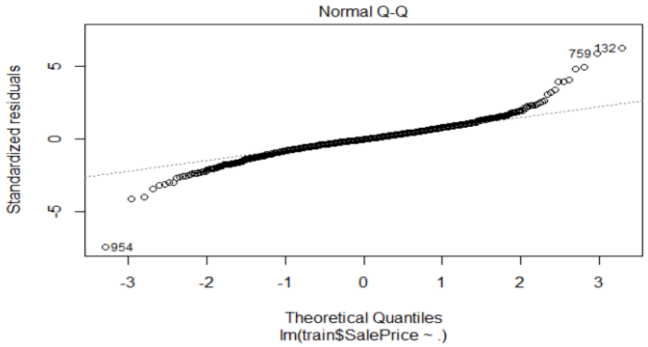
Below: output of a model with most variables included

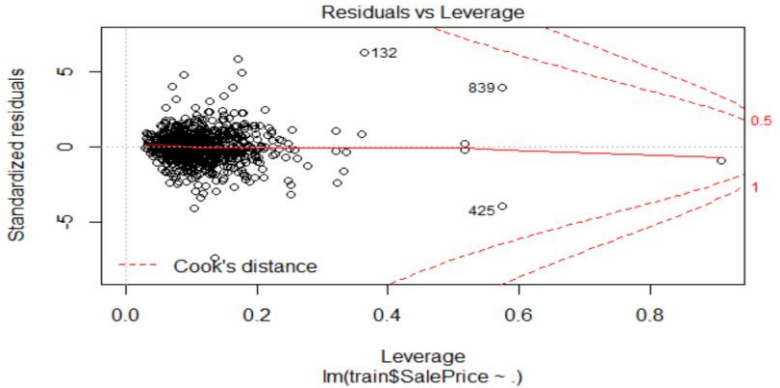
```
Residual standard error: 20700 on 879 degrees of freedom
  (58 observations deleted due to missingness)
Multiple R-squared:  0.9289,    Adjusted R-squared:  0.9199
F-statistic: 102.6 on 112 and 879 DF,  p-value: < 2.2e-16
```

Multiple Regression Model Plots









Lasso and Ridge Regression

Relying on the least squares estimates, lasso and ridge add a shrinkage penalty λ.

This reduces the B terms to exactly zero (lasso) or close to 0 (ridge).

- Lasso: helps with feature selection. Ridge: 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.
- We use cross-validation to find the right level of λ.

Lasso Regression

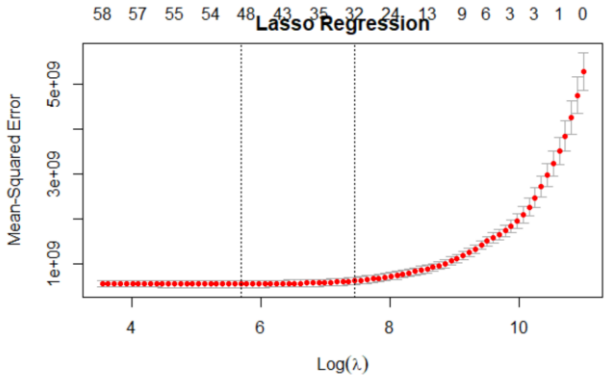
Split the train set using most variables

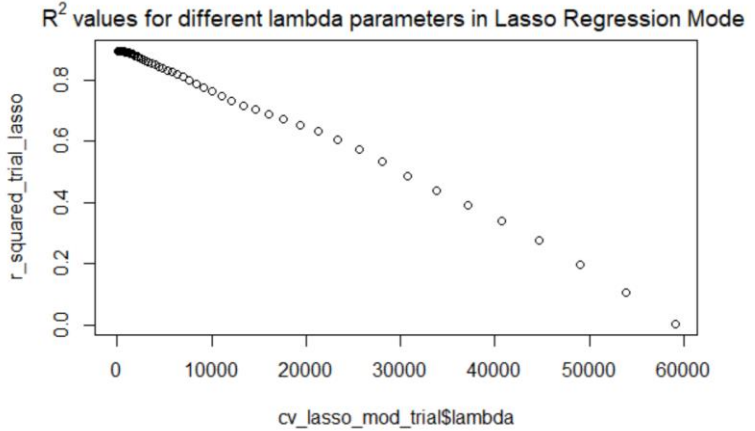
Run cross validation technique which standardizes the variables.

Cross validation produces a λ of 294.1
This is chosen as the lambda value for the smallest MSE

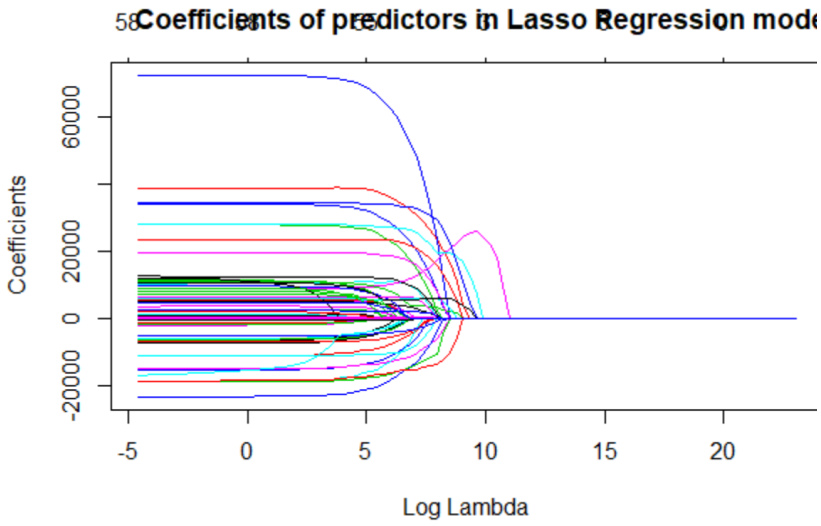
Then fit the model and predict values

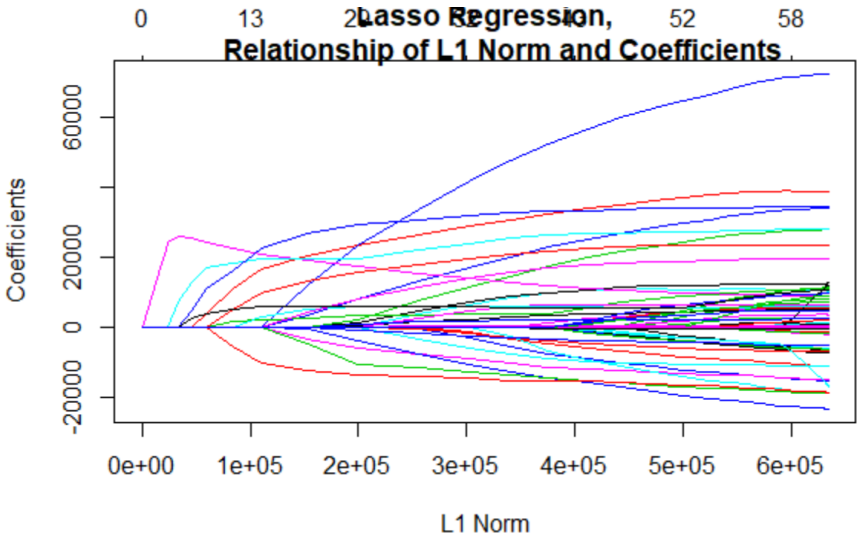
“Relaxed” lasso: Since lasso brings all variables towards 0, re-run lasso with only variables that have non-zero coefficients to get Bs for interpretation.





Lasso Regression (Cont.)





As λ increases, the coefficients shrink to exactly zero.
As the L1 Norm is Σ|B| increases, the coefficients increase

Ridge Regression Model

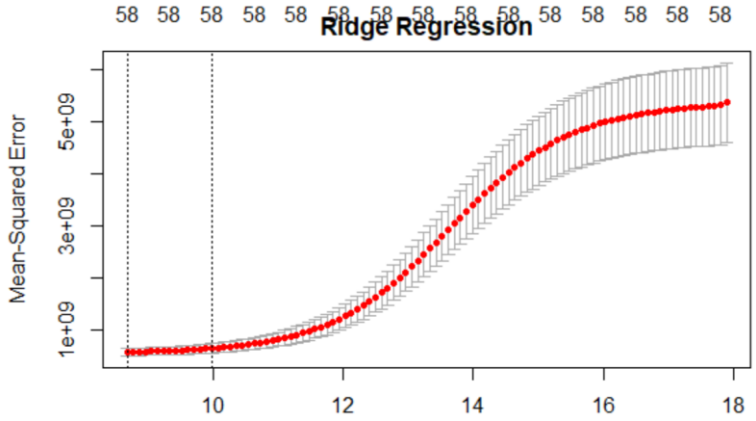
Split the train set using most variables

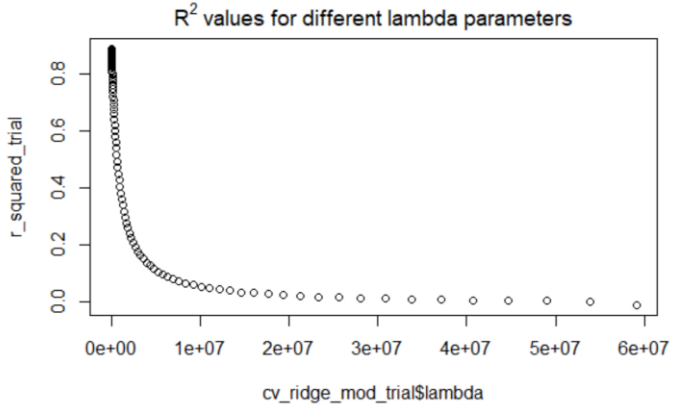
Run cross validation technique which standard
the variables.

Cross validation produces a λ of 5909.5
This is chosen as the lambda value for the smallest MSE

Then fit the model and predict values

The lowest point in the curve indicates the optimal lambda: the log value of lambda that best minimised the error in cross-validation.





Random forest

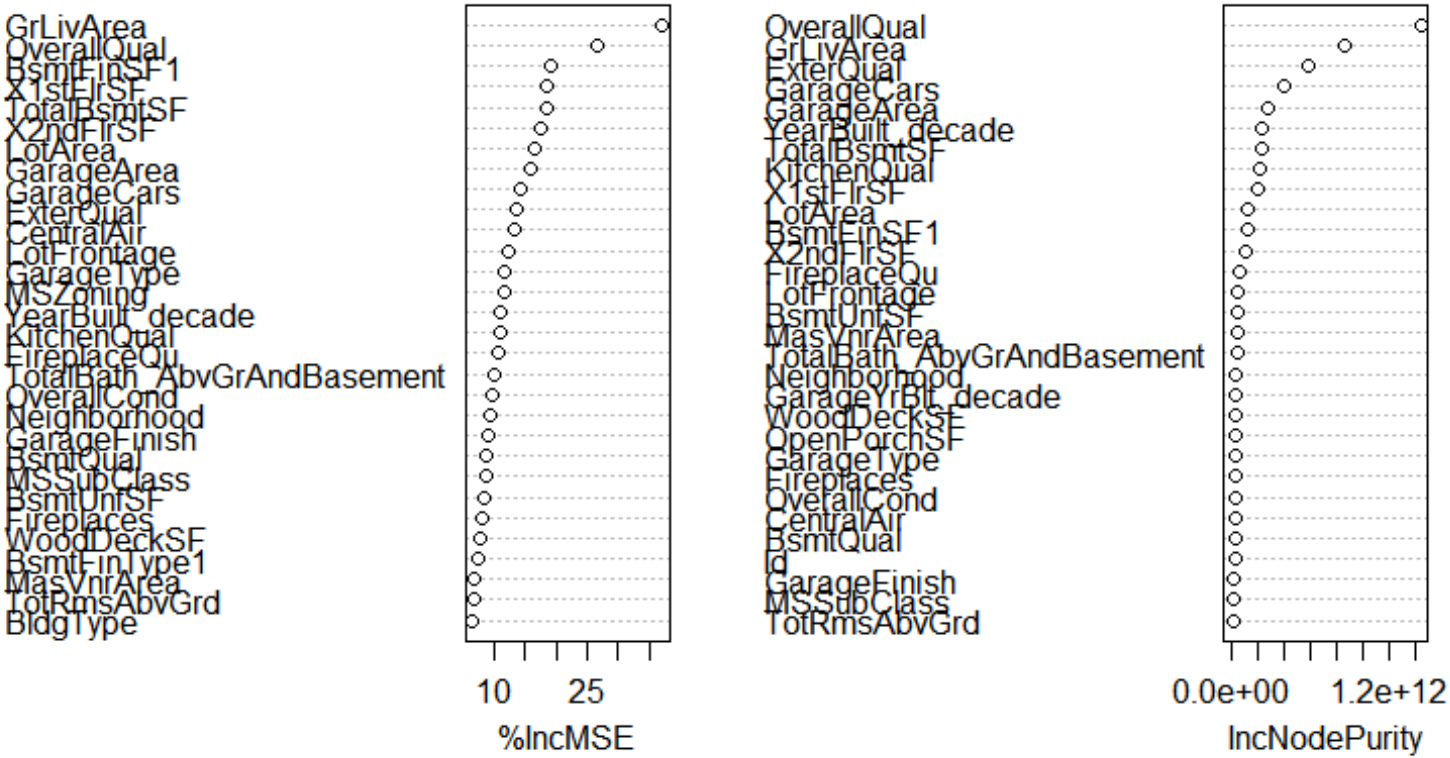
The shape of the model isn’t initially assumed (non-parametric)

Decision trees are used to see which observations fall into a specific node. Then we take the mean of observations in that node

- There could be a split at GrLivArea >= 1500 and GrLivArea < 1500

Random forest decorrelates the different trees by selecting a random predictor for the next level (usually in the set of $\sqrt{p}$)

Random forest model using 63 variables, Variable Importance Plot



Part IV: Model Validation

Deciding on the most important predictors

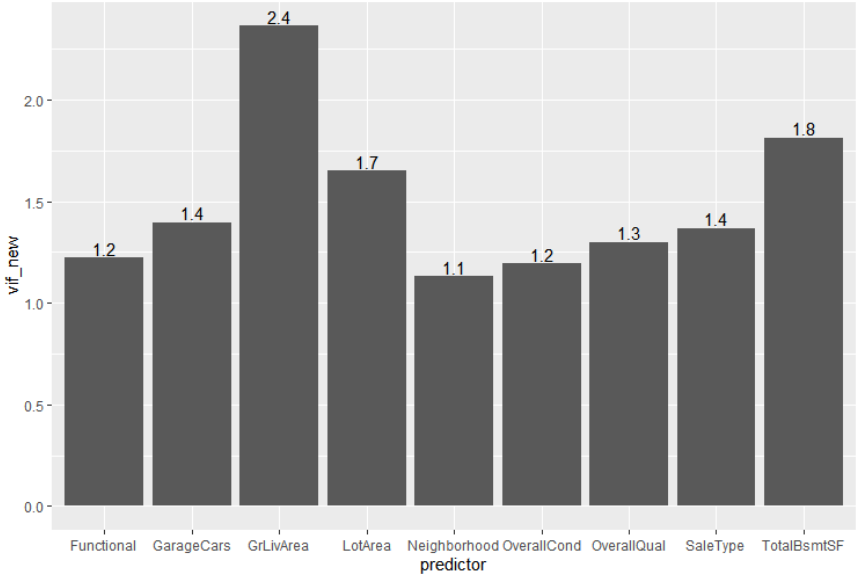
Relying on different models each of us ran, the group agreed on nine variables that accounted for an R^2 of 88.8%

- GrLivArea performed very well in RF not Lasso

There was an additional six variables a subset of us were interested in. The fifteen variable model had an R^2 of 89.8%

Although all the variables were significant in the fifteen variable model, we decided to go with the parsimonious one.

Multicollinearity was evaluated for the nine variables. We squared the generalized vif. All predictors were retained.



Part IV: Model Validation

- Use new data to check model and its predictive ability
  - Cross validation
  - Split data into k parts to reduce potential for over-fitting
- Compare results against test set to check predictive ability
  - See table below for evaluation criteria

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 models

GarageCars, Neighborhood, OverallCond
OverallQual, Functional, LotArea
SaleType, TotalBsmtSF, GrLivArea

9 variable model	Bias (using Mean)	Max Deviation	Mean Absolute Deviation	MSE	R^2
Multiple linear regression	-151	115,611	18,263	615,889,941	88.8%
Lasso regression	762	150,183	17,383	602,170,342	88.2%
Ridge regression	697	150,998	17,350	607,638,193	88.0%
Random forest	-11,298	268,938	23,579	765,709,896	85.9%

Generally, multiple linear regression, lasso and ridge regression performed similarly when using cross validation techniques.

Random forest performed worst on all measures but it helped us identify GrLivArea as very important in reducing MSE

Multiple linear regression is our model of choice.

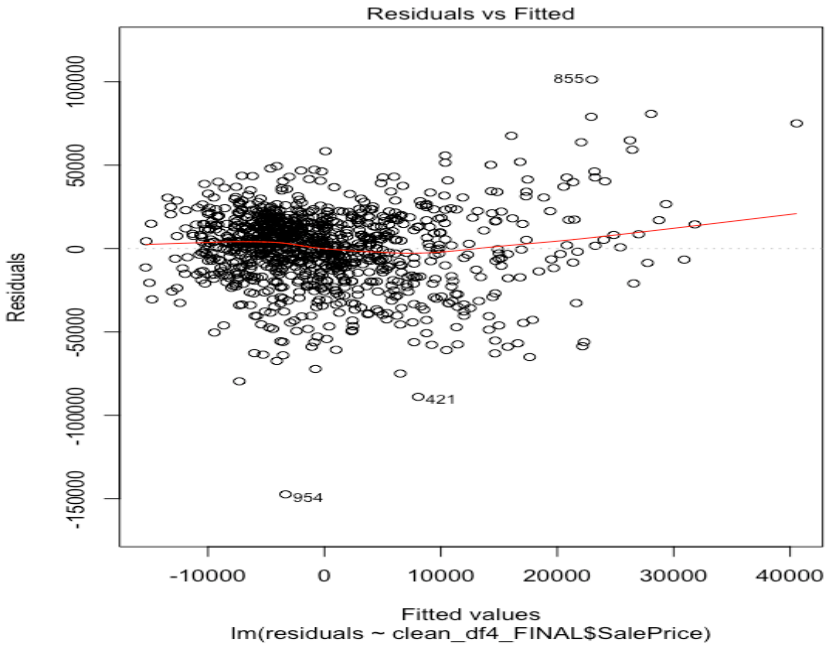
Final Multiple Model Regression Model: Summary

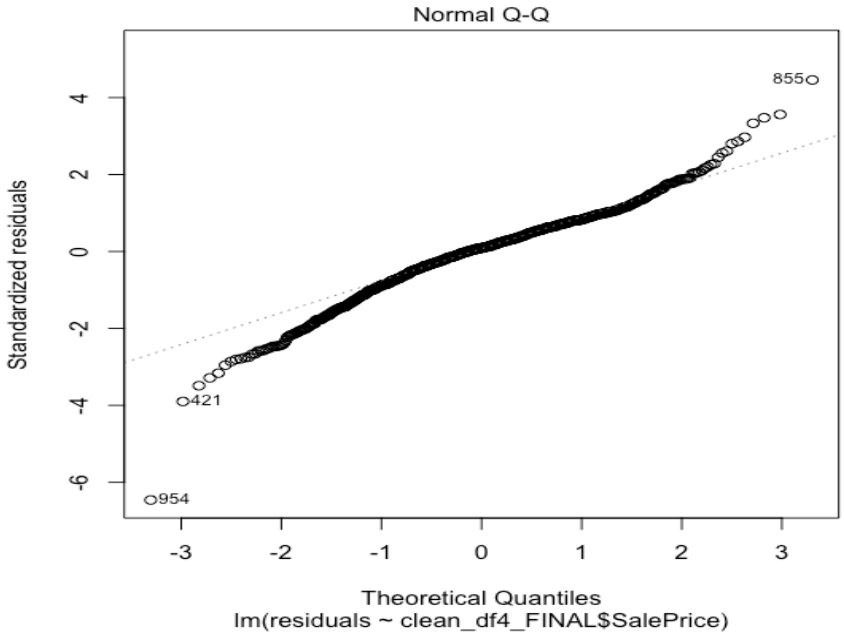
All nine variables in the final model were significant with p-values below 0.001

Feature coefficients (not shown) can be interpreted as the effect of a one unit increase in a variable, the SalePrice of a home will increase or decrease according to its coefficient, holding other variables constant

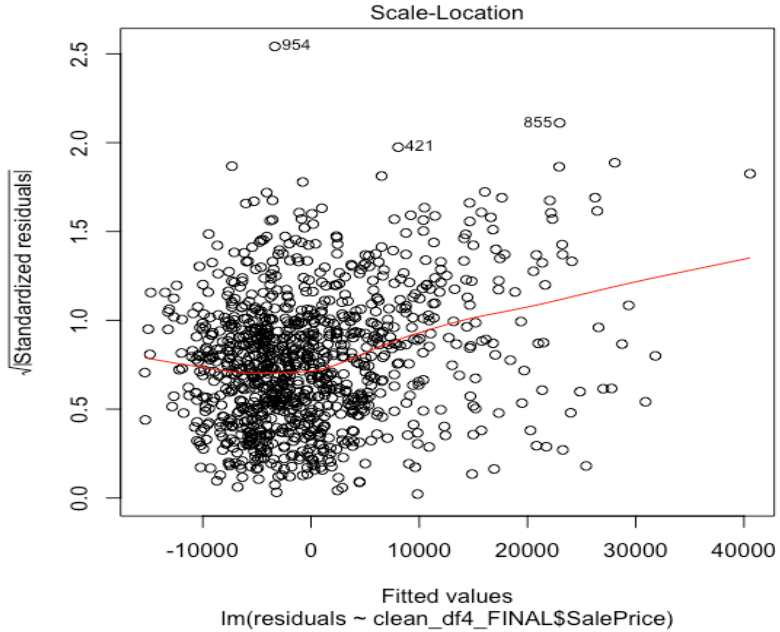
Feature	Df	SumSq	MeanSq	F value	P-value	Sig level
GarageCars	4	2.82E+12	7.04E+11	1150.9473	< 2.20E-16	***
GrLivArea	1	1.12E+12	1.12E+12	1831.4626	< 2.20E-16	***
OverallCond	7	1.65E+11	2.35E+10	38.4422	< 2.20E-16	***
OverallQual	9	6.07E+11	6.74E+10	110.1732	< 2.20E-16	***
LotArea	1	7.14E+10	7.14E+10	116.7603	< 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			

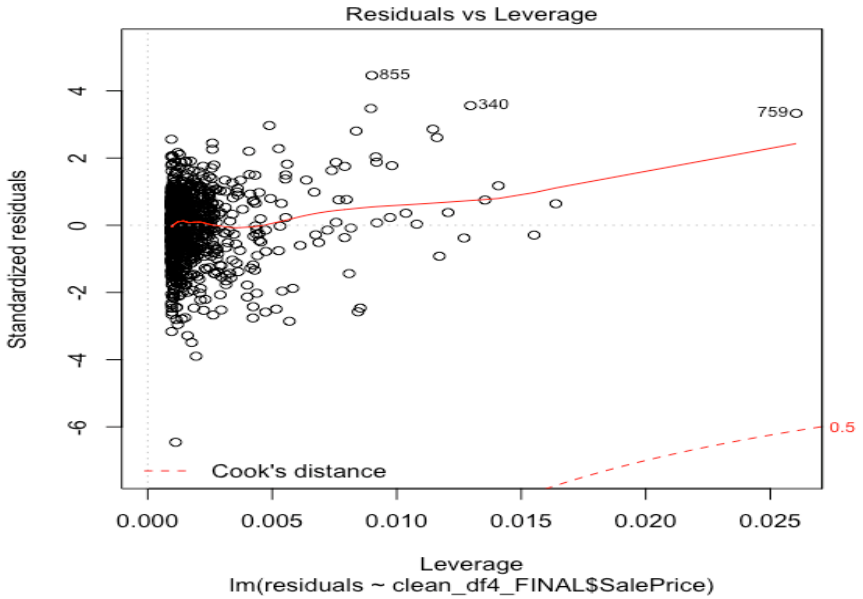
Plots of Final Multiple Linear Regression Model



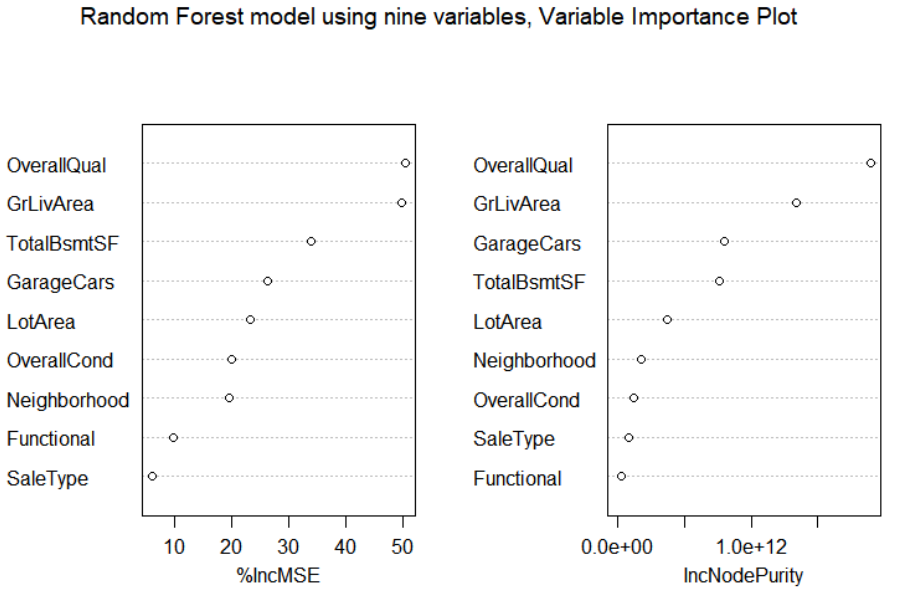


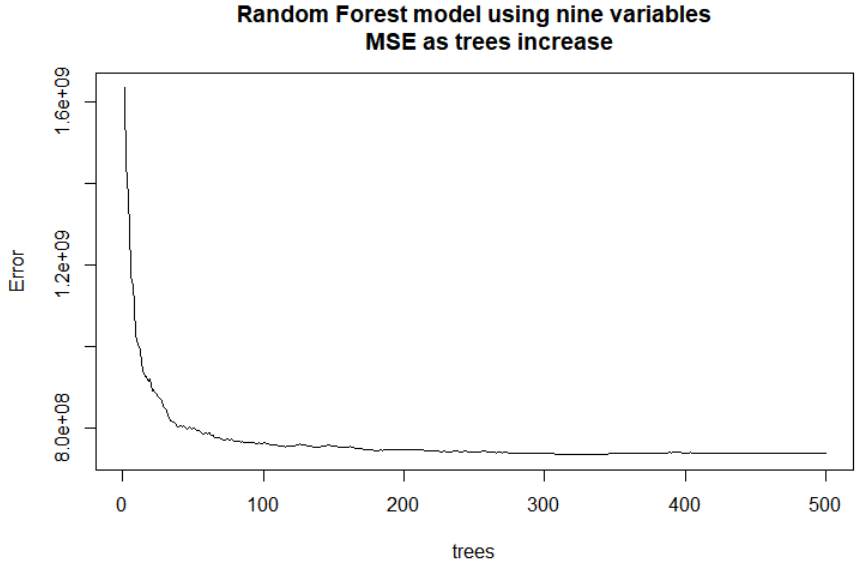
Plots of Final Multiple Linear Regression Model





Plots of Final Random Forest Model





Conclusion

The objective of this Kaggle competition was to build models to predict housing prices of different residences in Ames, IA.

The most important variables 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.