Real Estate Price Prediction and Quality Classification for Single Family Homes

Karim Bolous[1], Zach Gross[2], George Yuan[3]

Dec 13 2017

Executive summary

35% of all homebuyers in the U.S. are first time homeowners. The size of the entire U.S. real estate market is estimated at around $30 trillion. Every single person is involved in the real estate market in some capacity, so being able to model the market has the opportunity to affect every single person's life. Whether you're a seasoned real estate investor or a first time home buyer, it is important to understand trends in the housing market and being able to predict prices is an invaluable step towards that goal.

Using data from the Ames, Iowa Assessor's office, we were able construct models to predict home sale price and classify home quality. From this analysis, using Elastic Net, we were able to determine what aspects of a house genuinely increase its expected sale price and applied our model to determine whether homes in Ames on Zillow today are listed at a fair value. Equally importantly, we were also able to, using the Random Forest method, build a model to effectively classify the quality of homes, which, outside of price, is typically the most important factor under consideration by prospective homebuyers.

The price prediction model is a regression model selected from two candidate models using backward selection and elastic net trained on a subset of US housing data from Ames, Iowa. The classification model is selected from three candidate models using random forest, logistic classification, and neural net and classifies home quality into three levels - high, medium, low.

Goal of the study

We approached this study from the perspective of a recent college graduate seeking to purchase a home in the Ames, Iowa area in January 2011, the first month after end of our data. We wanted to solve the two big questions we thought such a person would have: is a given home fairly priced, and is it of high quality?

[1] Contact Karim at kgamil@wharton.upenn.edu
[2] Contact Zach at grossz@sas.upenn.edu
[3] Contact George at yuang@wharton.upenn.edu

To do this, we used the data available to generate a model to predict home sale price with the least possible prediction error, and created another model to classify home quality with the least misclassification error.

It was our goal in this analysis to build tools for new, inexperienced homebuyers in Ames to make data-driven choices for their purchase. These models would allow a shopper in Ames to do just that.

Limitations

One of the study’s main limitations is the limited scope of the data, which only focuses on homes in Ames, a small town north of Des Moines, Iowa. It would have been interesting to analyze a more exhaustive dataset from a geographical standpoint, as it would have allowed us to compare whether homes in different cities or states are characterized by similar predictor variables. Moreover, we believe it would have been interesting and insightful to expand the study to include homes sold before 2006 or after 2010, since we did not observe a significant change in average home price over the 5-year horizon. This also would have allowed us to analyze whether predictors of sale price varied over time, and whether the factors homeowners care most about change over time.

Description of Data

The raw dataset contained a total number of 82 variables. Because the raw dataset contained many redundant and obscure variables we removed 51 variables, with reasoning given below. For a description of the variables removed, see **Figure 1** in the Appendix. Also included in the appendix are plots used to justify some of our decisions. The variables kept are below:

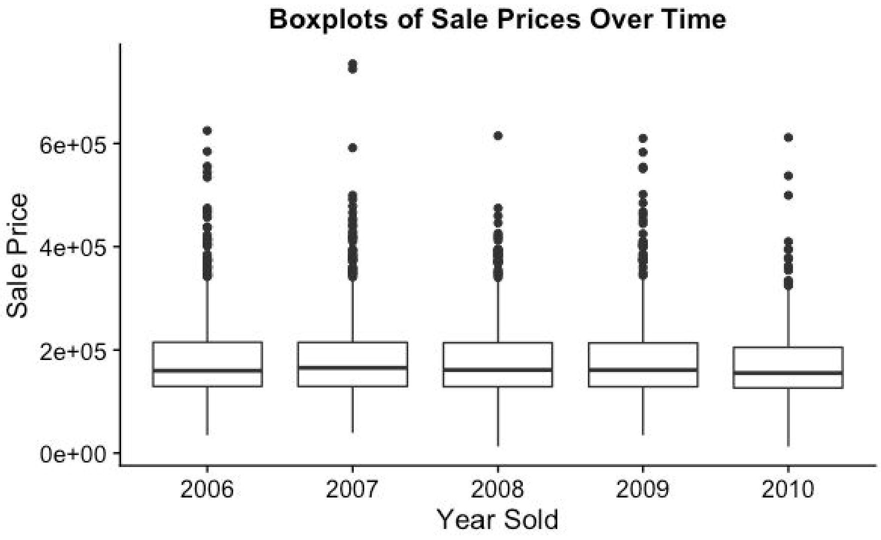
Variables such as order and PID were removed from the cleaned dataset since they were unique for each observation and could not be used for purposes of comparability or analysis. Many variables--such as lot shape, condition 1, condition 2, pool area, and pool quality--were removed from the original dataset because there was very little variability in the responses. For example, approximately 99.6% of the respondents stated that their houses did not have a pool, and most of the responses under condition 1 and condition 2 were normal.

Other variables--such as land contour, lot config, central.air, sale type, and misc.val--were removed due to the difficulty in interpreting these variables, as well as the fact that they may not be as likely to impact the price or quality of the house as much. Due to the large number of basement-related variables, we decided to keep TotalBsmt.SF in our clean dataset and remove all other basement variables (such as Bsmt.Cond, Bsmt.exposure, BsmtFin.Type1, BsmtFin.Type2, BsmtFin.SF.1, BsmtFin.SF.2 and Bsmt.Unf.SF). This was because we believed that the total area of the basement was more likely to serve as a better indicator for predicting home prices and quality, compared to the type of rating of the basement.

Following the same line of logic, we decided to remove all garage-related variables (such as garage.finish, garage.cars, and garage.cond) except garage area, since the collinearity between garage area and other garage-related variables was very high. Lastly, variables (such as fireplace.qu and alley) that included a large number of N/A's or missing responses were removed.

Analysis

Exploratory Data Analysis



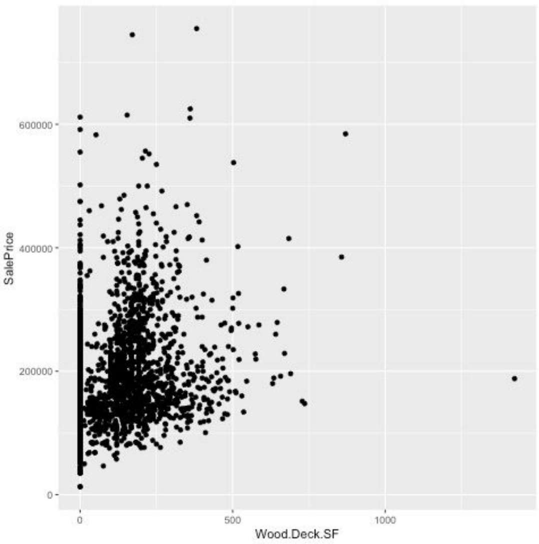
Because the data was collected between 2006 and 2010, we suspected that the 2008 recession had an impact on homes sold after 2008 and decided to plot a boxplot of year sold versus sale price. Surprisingly though, the box plot shows no noticeable change in the mean house price over the four years.

Since we were highly skeptical that the recession had no significant impact on home prices in Ames, Iowa, we researched the least impacted cities in the U.S. by the 2008 recession to corroborate this observation. According to a CNBC article, Des Moines (a city 30 miles from Iowa) was ranked as the 13th least impacted city by the recession in the U.S., corroborating the observation from the boxplot.

In order to analyze the impact of the exterior covering and quality on the sale price of the house, a boxplot of exterior.1st was plotted and was filled with exter.quality (quality of the exterior covering), which can be found under **Figure 4** in the Appendix. The results showed two main observations. Not surprisingly, homes that had a higher exterior quality on average had a higher sale price. Moreover, the material of the exterior covering had somewhat of an impact on sale price--homes with an exterior of cement board and vinyl siding on average had the highest sale price.

Similar to the results observed from the quality of the home's exterior covering, homes that had larger basements (in terms of both area and height) had a higher sale price on average. Interestingly, Bsmt.Qual (height of the basement) seemed to have a larger impact on the sale price of the house than did Total.Bsmt.SF (area of the basement), suggesting that home buyers seemed to care more about the basement's height than its area. This plot can be found under **Figure 2** in the Appendix. Variables that were found to have little to no impact on a home's sale price were wood deck area and type of sale, which makes sense since the type of contract or warranty on the home has little to no impact on the value of the home.

Our EDA also revealed that the deck- and porch-area related variables were weirdly distributed in the data. Take for example Wood Deck area:



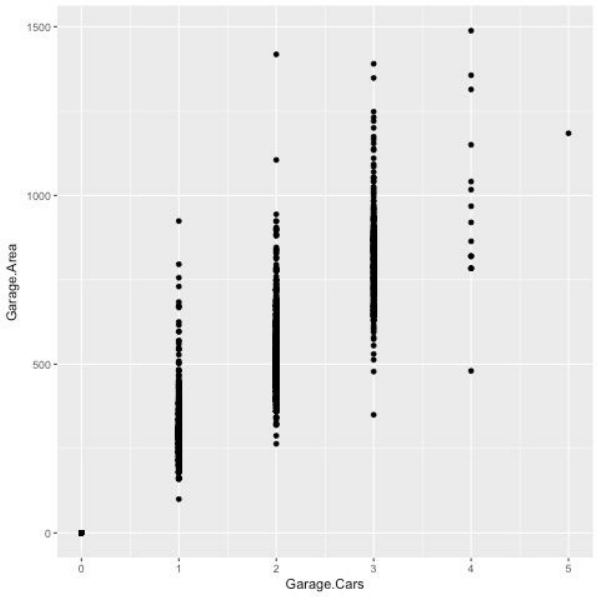
There appears to be an interesting relationship between Sales Price and Wood Deck area, but there are also a number of observations with no wood decks (including many with high sales prices). Because of this, we recategorized the following five variables:

- Wood.Deck.SF (wood deck area)
- Open.Porch.SF (open porch area)
- Enclosed.Porch (enclosed porch area)
- Screen.Porch (screen porch area)
- X3Ssn.Porch (3-season porch area)

Instead of being continuous variables measuring the area of a deck or porch, the variables were changed to dummy variables indicating whether the house had such a feature.

Finally, as mentioned in our data description section, the Ames dataset was overly rich in some areas for our purpose. So many variables were given for areas of the house such as the garage and basement that leaving too many in would raise questions of collinearity. This can be seen in the plot below:

Unsurprisingly, there is a close relationship between the size of a garage and the amount of cars it can fit. Thus, we kept the variables in data-rich areas that we thought would give the most variation.

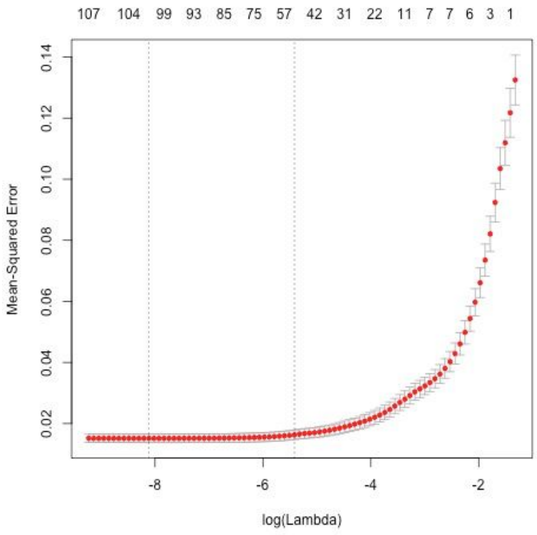


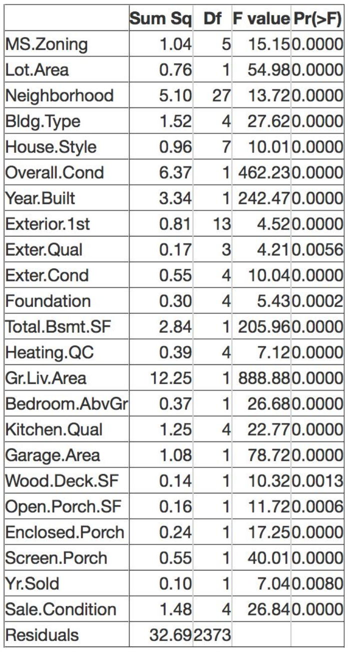
Price Prediction

The goal of this part of our analysis was to build a model to value home prices in the Ames, Iowa area in order to do determine if listings on the internet were fairly valued. To do this, we attempted two model selection methods: Elastic Net and Backward Selection.

Elastic Net

Due to the large number of variables in the dataset, Elastic Net was used to identify relevant factors that impacted the sale price of a home. Alpha was set to .99 and the number of folds chosen for cross validation was 10. The highest lambda within one standard error of the minimum (as seen in the plot to the left) was selected as the threshold for our predictors.





	Sum Sq	Df	F value	Pr(>F)
MS.Zoning	1.04	5	15.15	0.0000
Lot.Area	0.76	1	54.98	0.0000
Neighborhood	5.10	27	13.72	0.0000
Bldg.Type	1.52	4	27.62	0.0000
House.Style	0.96	7	10.01	0.0000
Overall.Cond	6.37	1	462.23	0.0000
Year.Built	3.34	1	242.47	0.0000
Exterior.1st	0.81	13	4.52	0.0000
Exter.Qual	0.17	3	4.21	0.0056
Exter.Cond	0.55	4	10.04	0.0000
Foundation	0.30	4	5.43	0.0002
Total.Bsmt.SF	2.84	1	205.96	0.0000
Heating.QC	0.39	4	7.12	0.0000
Gr.Liv.Area	12.25	1	888.88	0.0000
Bedroom.AbvGr	0.37	1	26.68	0.0000
Kitchen.Qual	1.25	4	22.77	0.0000
Garage.Area	1.08	1	78.72	0.0000
Wood.Deck.SF	0.14	1	10.32	0.0013
Open.Porch.SF	0.16	1	11.72	0.0006
Enclosed.Porch	0.24	1	17.25	0.0000
Screen.Porch	0.55	1	40.01	0.0000
Yr.Sold	0.10	1	7.04	0.0080
Sale.Condition	1.48	4	26.84	0.0000
Residuals	32.69	2373		

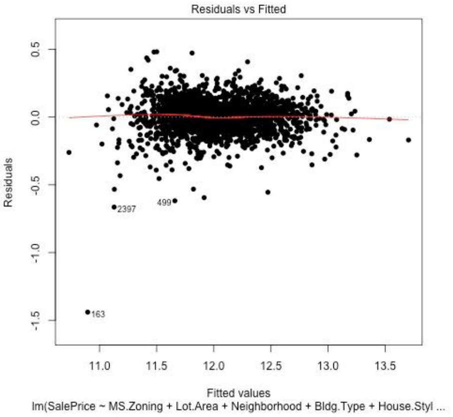
Elastic Net under these parameters returned 25 non-zero betas. A linear regression was then fit on log(SalePrice) using these predictors and a type 2 Anova test was run to determine the variables' significance. Predictors not significant at the .01 level were removed one at a time based on which had the highest p-value. This led to the removal of two variables: Exterior.2nd and Year.Remod.Add. Our final Elastic Net model thus had 23 predictors (see image above).

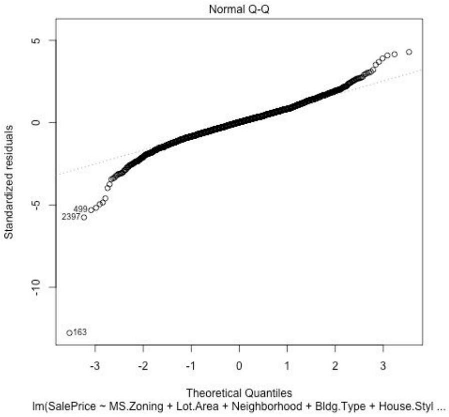
Backward Selection

The second model selection method we used to determine sales price was backward selection through the 'regsubsets' function in the 'leaps' package. We chose the set of variables that minimized Cp (prediction error) and fit them with OLS. Like we did with the Elastic Net model, we checked that each predictor was significant at the .01 level using a Type 2 Anova test. This gave us a model identical to the one we found through Elastic Net.

Model Diagnostics

We then checked to see if the model's residuals were normally distributed, independent, and uncorrelated using a Residual Scatter plot and a Q-Q plot.

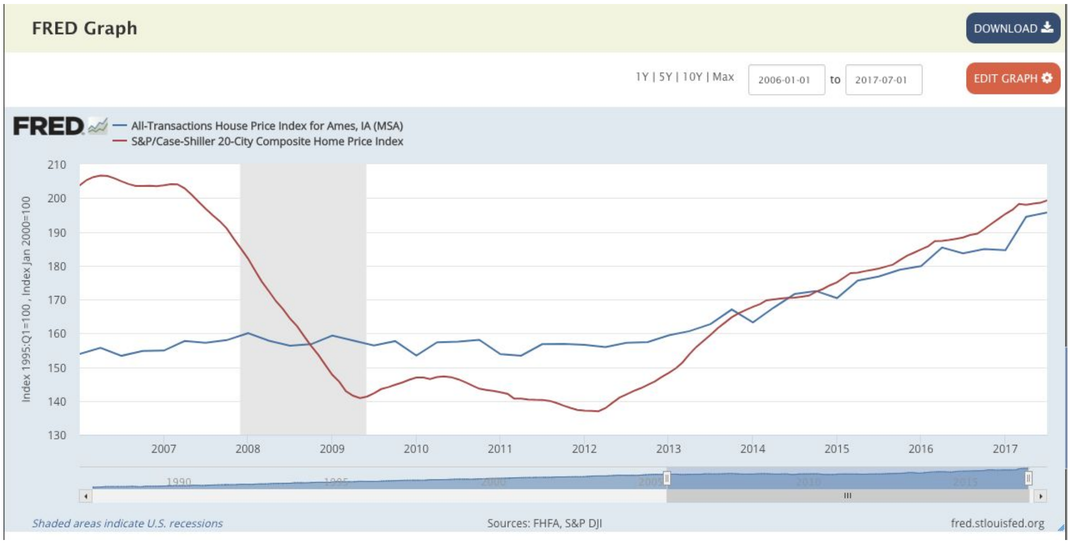




The results from the graphs were in line with our model assumptions.

Application

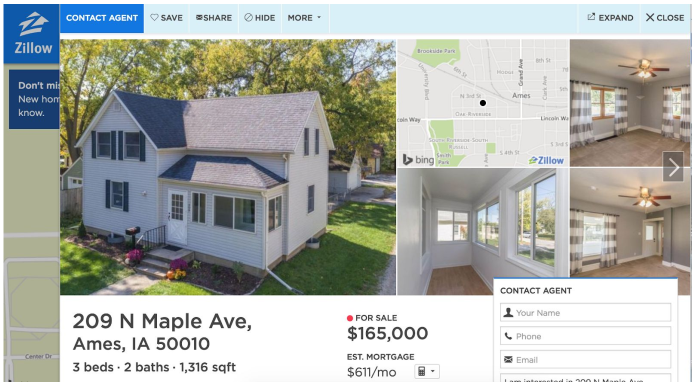
The purpose of our model was to determine if home prices listed on Zillow were overpriced. However, our data only covers house sales from 2006-2010, while Zillow only lists the asking price of houses currently on the market. To rectify this, we rescaled the prices on Zillow using the Ames Housing Price Index from the U.S. Federal Housing Finance Agency to get a rough estimate of what they would have been in January 2011. The below graph compares the Ames Housing Price Index with the Case-Shiller Index, which indexes housing prices in the entire country.



The ratio of housing prices in Ames in July 2017 to prices in Ames in January 2011 was approximately 0.79, so we multiplied the prices listed on Zillow by that fraction to get our estimate for what their 2011 Zillow listing price would have been.

We evaluated the price of two houses in a neighborhood near Iowa State University, a location we expect would have a high student and recently-graduated population.

House 1:

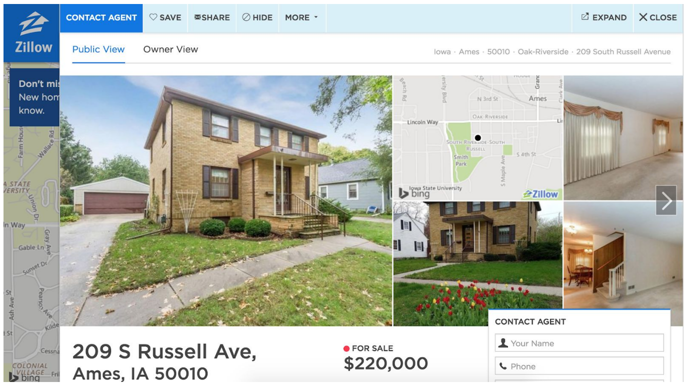


209 N Maple Ave,
Ames, IA 50010
3 beds · 2 baths · 1,316 sqft

FOR SALE
$165,000
EST. MORTGAGE
$611/mo

Your Name
Phone
Email
I am interested in 209 N Maple Ave

House 2:

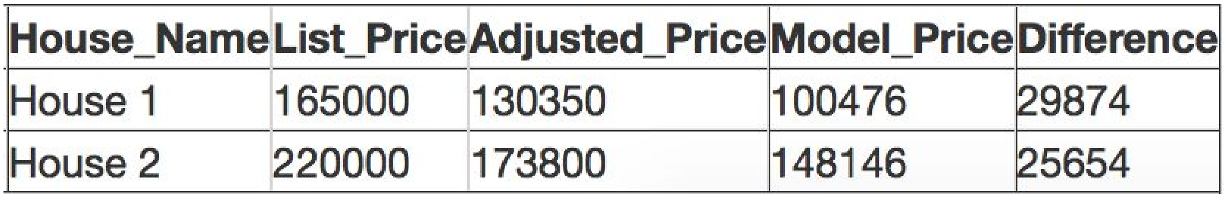


209 S Russell Ave,
Ames, IA 50010

FOR SALE
$220,000

Your Name
Phone

Each house was listed with a parcel number, an identification code that allowed us to look up the property in the same Ames Assessors' office from which our original dataset was sourced. We inputted the relevant data on these homes from the Ames Assessors' office into our model and received the following results.



House_Name	List_Price	Adjusted_Price	Model_Price	Difference
House 1	165000	130350	100476	29874
House 2	220000	173800	148146	25654

Our model found that both of the homes under consideration on Zillow were overpriced by over $25,000, suggesting that we should look elsewhere to find fairly-valued housing.

House Quality Classification

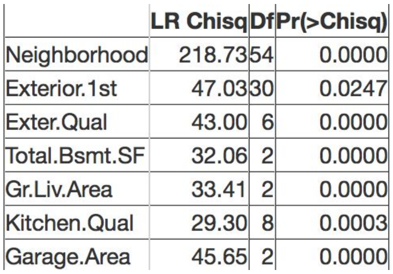
After modeling SalePrice, we attempted to model whether a house was of high, medium, or low quality. This data was included in the dataset as an integer column taking on values between 2 and 10 and we reconfigured it for modeling purposes to take on the value low if it was between 2 and 4, medium if it was between 5 and 7, and high if it was between 8 and 10. The three approaches we took were multinomial logistic regression, Random Forest, and Neural Net. Overall, we selected Random Forest because it yielded the lowest misclassification error.

Logistic Regression

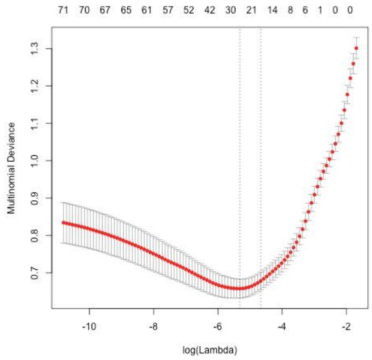
Because overall quality was reclassified earlier as a factor with 3 levels (low, medium and high), cv.glmnet was run using the multinomial family. The plot resulting from this code is displayed below.

The elastic net yielded 9 different variables using fit.1se as the criterion.

A logistic regression was then fit using the 9 variables and an Anova test was carried out on the regression- any variables that were not significant at the 0.05 level were removed. It makes intuitive sense that variables like MS.Zoning, which were removed, do not impact the overall quality of the house. However, factors in the final logistic model, like Kitchen Quality and Exterior Quality, are more likely to have a greater impact on the overall quality of the house. The final multinomial logistic model had 7 factors.



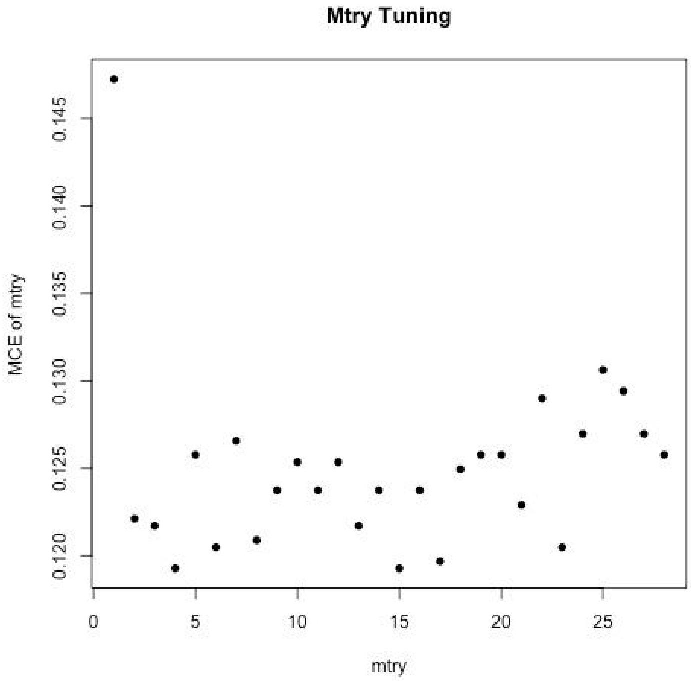
	LR Chisq	Df	Pr(>Chisq)
Neighborhood	218.73	54	0.0000
Exterior.1st	47.03	30	0.0247
Exter.Qual	43.00	6	0.0000
Total.Bsmt.SF	32.06	2	0.0000
Gr.Liv.Area	33.41	2	0.0000
Kitchen.Qual	29.30	8	0.0003
Garage.Area	45.65	2	0.0000



The misclassification error for the model was 0.186.

Random Forest

We build a random forest model using 100 trees. The number of subset variables that are used for each tree is set to be a tuning parameter that we found to be optimal at 4. Optimality is evaluated on the basis of misclassification error. The misclassification error of our random forest classification was found to be 0.119.



Neural Net

Our final classification model was a neural net. It was trained on 1965 observations and tested on 500 others. All columns that were retained from the initial data cleaning process were included. Categorical variables were converted into dummy variables for each factor level using the model matrix function. The only data that was excluded were factor levels which could not be scaled - MS.ZoningI, FoundationSlab, and SaleCondition.Partial. These levels were so infrequently found in the data that the sample data did not include any observations where they were present. This meant that when these columns were scaled they would return ‘NaN’ values, which the neural net would not accept as input. After scaling, the data was fit onto a neural net with 55 neurons and 1 layer, and the function outputted the probability that a given house was of low, medium, or high quality. Overall, the misclassification error for this neural net on the testing data was 0.224, which was relatively low, but still higher than Random Forests’ and Logistic’s MCE.

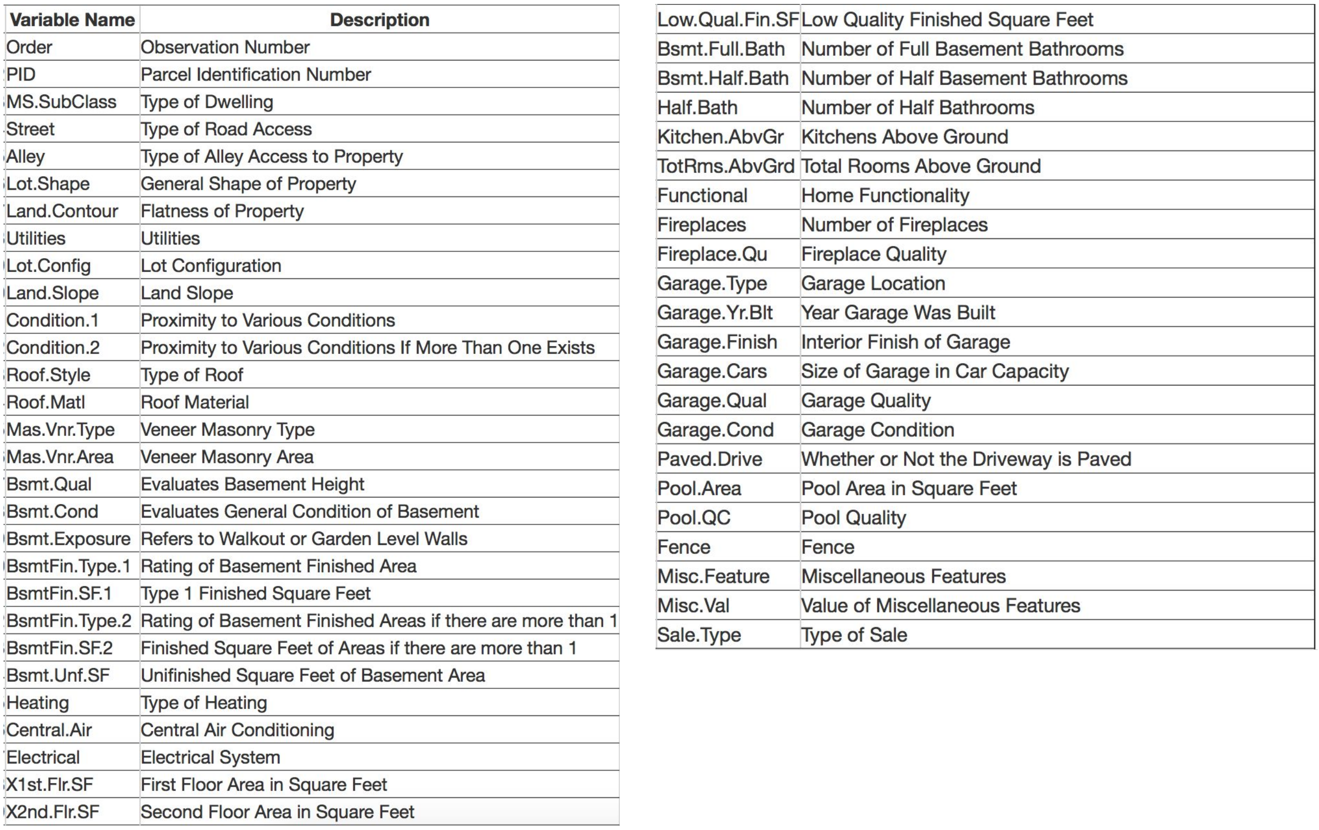
Conclusion

Acknowledgements

Special thanks to Professor Linda Zhao for making this project possible, and for providing us with the numerous tools that we used to make this possible. Additional thanks to the Ames Assessor’s office for the detailed and meticulously collected data, and to Dean de Cock at Truman State University for compiling our original dataset and making it available.

Appendix

Figure 1: Removed Variables



Variable Name	Description
Order	Observation Number
PID	Parcel Identification Number
MS.SubClass	Type of Dwelling
Street	Type of Road Access
Alley	Type of Alley Access to Property
Lot.Shape	General Shape of Property
Land.Contour	Flatness of Property
Utilities	Utilities
Lot.Config	Lot Configuration
Land.Slope	Land Slope
Condition.1	Proximity to Various Conditions
Condition.2	Proximity to Various Conditions If More Than One Exists
Roof.Style	Type of Roof
Roof.Matl	Roof Material
Mas.Vnr.Type	Veneer Masonry Type
Mas.Vnr.Area	Veneer Masonry Area
Bsmt.Qual	Evaluates Basement Height
Bsmt.Cond	Evaluates General Condition of Basement
Bsmt.Exposure	Refers to Walkout or Garden Level Walls
BsmtFin.Type.1	Rating of Basement Finished Area
BsmtFin.SF.1	Type 1 Finished Square Feet
BsmtFin.Type.2	Rating of Basement Finished Areas if there are more than 1
BsmtFin.SF.2	Finished Square Feet of Areas if there are more than 1
Bsmt.Unf.SF	Unifinished Square Feet of Basement Area
Heating	Type of Heating
Central.Air	Central Air Conditioning
Electrical	Electrical System
X1st.Flr.SF	First Floor Area in Square Feet
X2nd.Flr.SF	Second Floor Area in Square Feet
Low.Qual.Fin.SF	Low Quality Finished Square Feet
Bsmt.Full.Bath	Number of Full Basement Bathrooms
Bsmt.Half.Bath	Number of Half Basement Bathrooms
Half.Bath	Number of Half Bathrooms
Kitchen.AbvGr	Kitchens Above Ground
TotRms.AbvGrd	Total Rooms Above Ground
Functional	Home Functionality
Fireplaces	Number of Fireplaces
Fireplace.Qu	Fireplace Quality
Garage.Type	Garage Location
Garage.Yr.Blt	Year Garage Was Built
Garage.Finish	Interior Finish of Garage
Garage.Cars	Size of Garage in Car Capacity
Garage.Qual	Garage Quality
Garage.Cond	Garage Condition
Paved.Drive	Whether or Not the Driveway is Paved
Pool.Area	Pool Area in Square Feet
Pool.QC	Pool Quality
Fence	Fence
Misc.Feature	Miscellaneous Features
Misc.Val	Value of Miscellaneous Features
Sale.Type	Type of Sale

Figure 2: Basement Square Footage by Sales Price

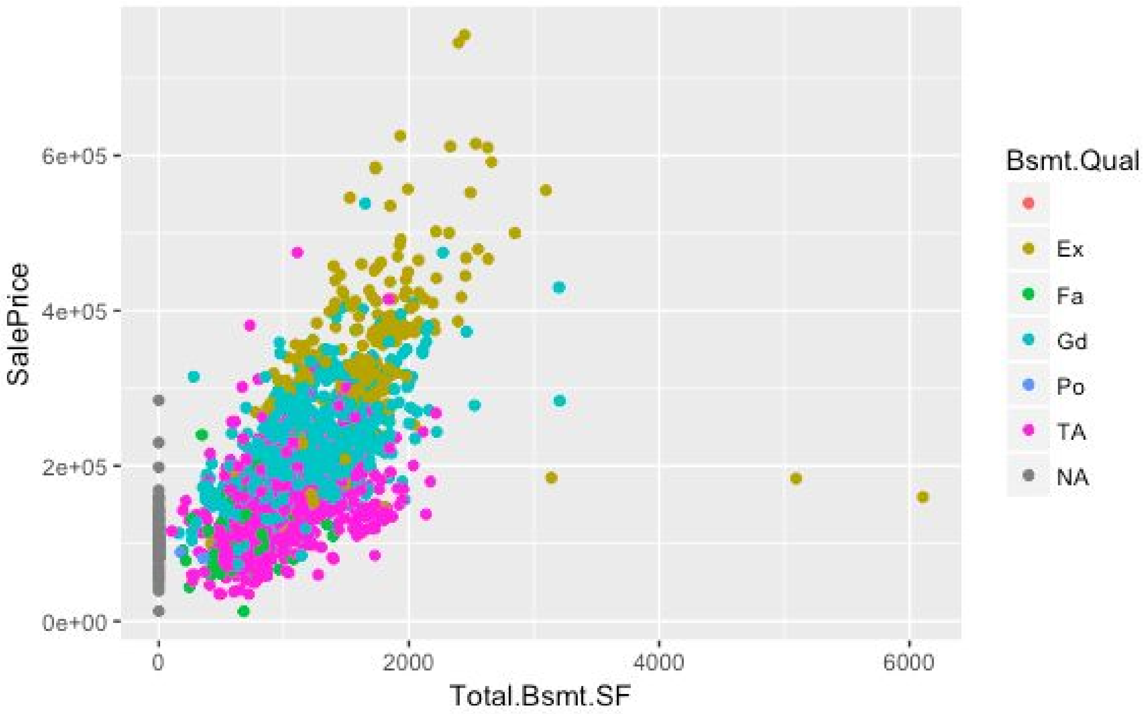


Figure 3: Random Forest ROC Output

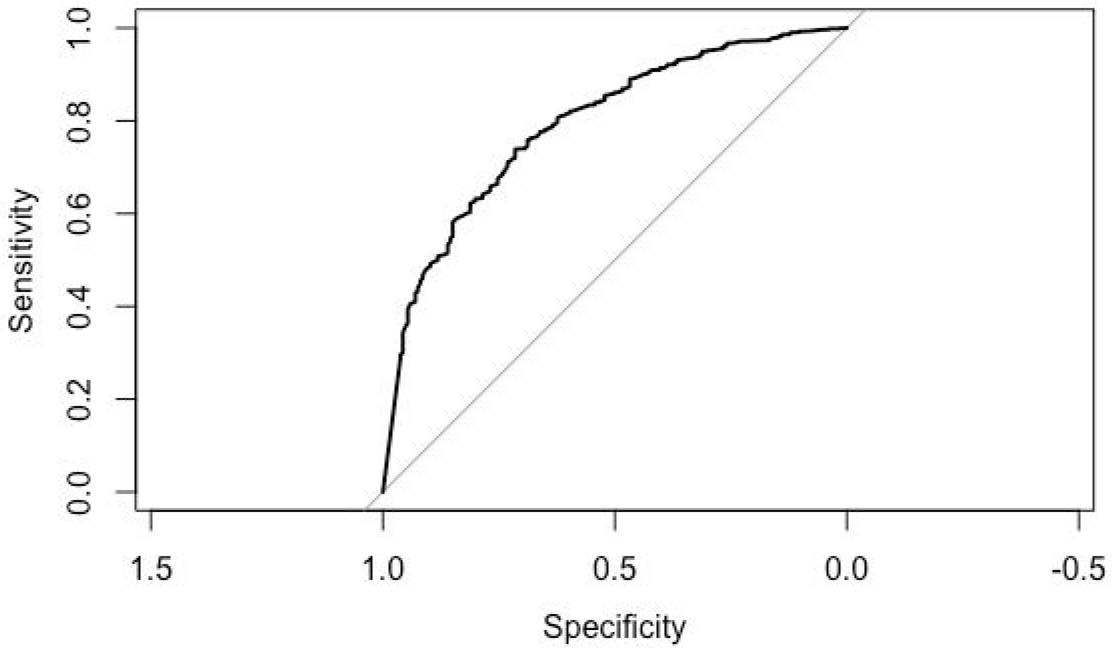
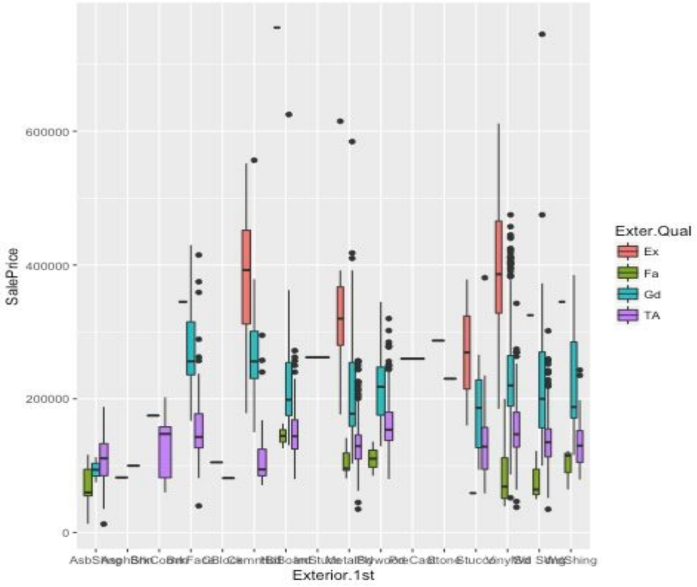


Figure 4: Exterior material vs. Sales Price by Exterior Quality


SalePrice
Exterior.1st
AsbShng
AsphShn
BrkComm
BrkFace
CBlock
CemntBd
HdBoard
ImStucc
MetalSd
Plywood
PreCast
Stone
Stucco
VinylSd
Wd Sdng
WdShing
Exter.Qual
Ex
Fa
Gd
TA