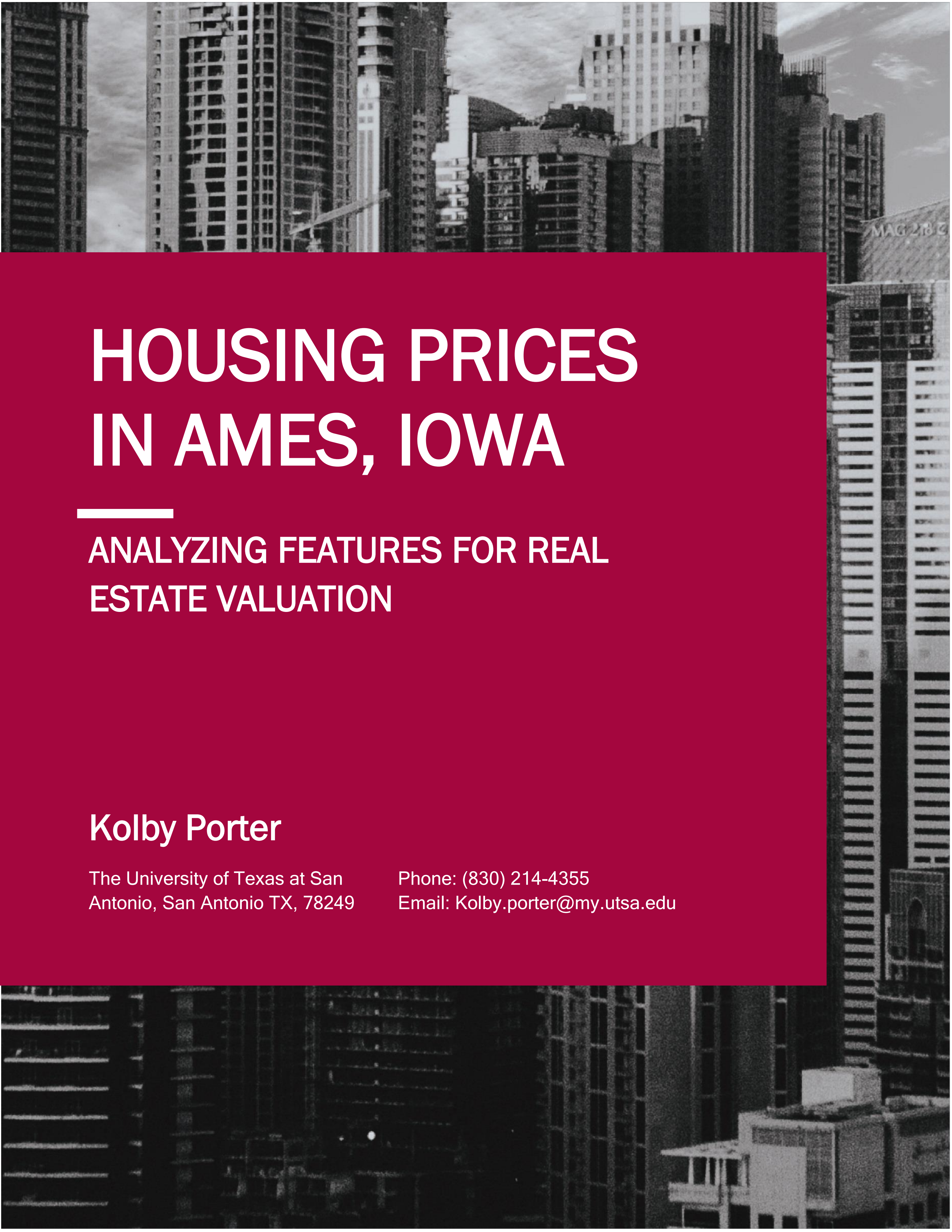


HOUSING PRICES IN AMES, IOWA

ANALYZING FEATURES FOR REAL ESTATE VALUATION

Kolby Porter

The University of Texas at San
Antonio, San Antonio TX, 78249

Phone: (830) 214-4355
Email: Kolby.porter@my.utsa.edu

Sale Price Valuation – A Prediction of Critical Features

Introduction

The residential real estate market has been a centerpiece of discussion in recent years following the COVID-19 pandemic. In levels never observed before, much of the American workforce found themselves working from home to limit COVID exposure. With interest rates rising to combat inflation, the housing market has proceeded in kind. Residential real estate is at an all-time high. For both buyers and sellers in today's market, it's paramount to understand what exactly dictates the going price for a given dwelling. To understand that I used a multitude of characteristics that could describe a particular home and analyzed those to determine the most impactful. My analysis used the House Prices - Advanced Regression Techniques dataset provided by Anna Montoya and Data Canary (2016) from Kaggle.

BACKGROUND

The data mentioned above contains the characteristics – hereafter referred to as features and or explanatory variables – of roughly three-thousand dwellings in Ames, Iowa. The dwellings included in the study were sold between 2006 and 2010. This data goes above and beyond to catalogue many relevant characteristics of each home. Seventy-nine explanatory variables were included in said data; some described the neighborhood a home resided in or immediate road access. For the most part though, variables described physical features, such as roof type, basement height, and number of bathrooms.

Objective

My goal was to identify the characteristics that most significantly impact home prices and to validate these findings by predicting the prices of homes with similar attributes using the data from this study. Key characteristics will be the variables that have the most impact on sale price and drive my predictions. To accomplish this, predictive modeling was conducted using R. Multiple models were developed and trained on one dataset, and then tested on a separate dataset to predict sale prices. This process revealed several characteristics with minimal predictive value, leading to their exclusion to minimize noise and enhance model accuracy.

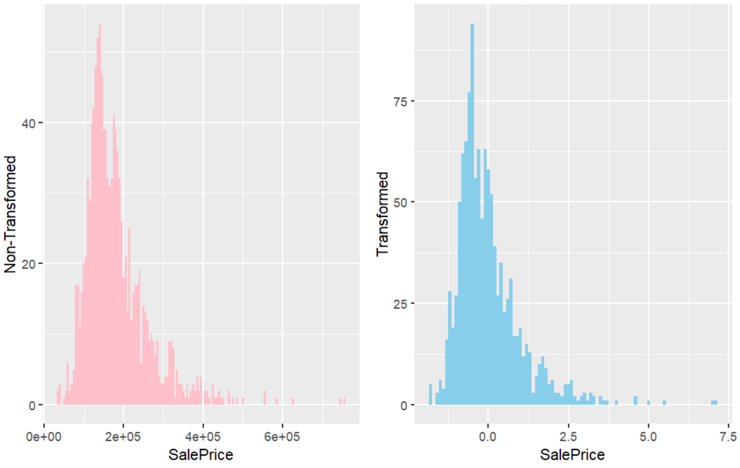
Methods

In analyzing the data, the first step was to clean up the given observations and variables. Most characteristics that were not present in each variable were initially coded as NA. These NAs were subsequently converted to 0 to retain as much data as possible for analysis.

Several variables from this data represented different categorical aspects of a single characteristic. An example of that being “RoofStyle”, this feature contained multiple options that corresponded with the style of roof that the given dwelling was built with. That variable, and the variables like it, were one-hot coded to become binary indicators. The new variable became “RoofStyle_Gable” and held a 1 for gabled roofs and 0 for all others. This allowed me to retain as much data as possible and analyze certain characteristics in a meaningful way.

After one-hot coding, the set of predictors grew by quite a bit. Predictors that had zero or near zero variance were removed to reduce dimensionality and avoid potential noise. Variables with no variance provided no insight into sale price, as their lack of variation prevented the establishment of any correlation. By doing this, most of the characteristics that had no predictive power were removed.

After cleaning the data, a minimal level of preprocessing was done to center and scale for uniform analysis. Transformation methods such as “Box-Cox” and “Yeo Johnson” were avoided as the data was relatively normal to begin with and predicting real sale prices for the test data would’ve been very cumbersome. The distribution of the data before and after scaling and centering is shown by the figure to the right.



Several models were tested, but ultimately an adaptive variable neural network model was used for my prediction. Support-vector machine (SVM), partial least squares (PLS) and a multivariate adaptive regression splines (MARS) models were also considered for this analysis.

While the neural network was more computationally complex and needed extra effort to reduce over fitting, it predicted data more accurately than the rest. A support vector machine model proved to be statistically equivalent to the neural network when data was transformed using the Box-Cox method. The models trained and tested performed to the following levels without said transformation.

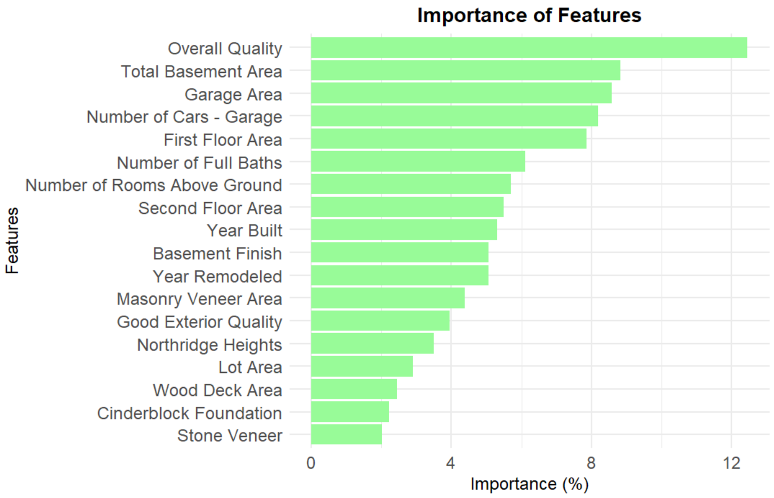
Model Performance			
	RMSE	R squared	MAE
SVM	0.4566	0.8087	0.2341
PLS	0.4687	0.7833	0.2678
NNet	0.3692	0.8724	0.2217
MARS	0.5999	0.6951	0.2908

The neural network model, or NNet, as pictured above, outperformed the other models in RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and the R squared metric.

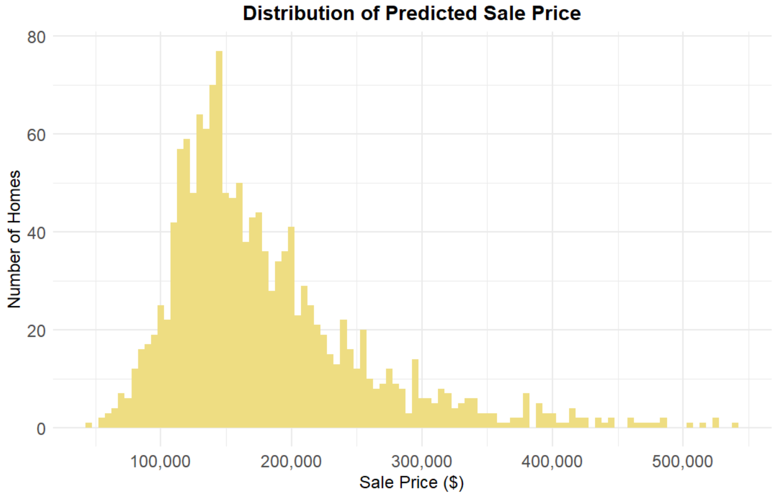
RMSE is an error metric that measures the standard deviation of the residuals (prediction errors). It provides an indication of how well the model's sale prices line up with the actual data. MAE measures the average value of the errors in a set of predictions. It calculates the average absolute differences between the predicted sale price and the actual sale prices. In short, R squared measures how well the variance in our sale price is explained by the given model

Results

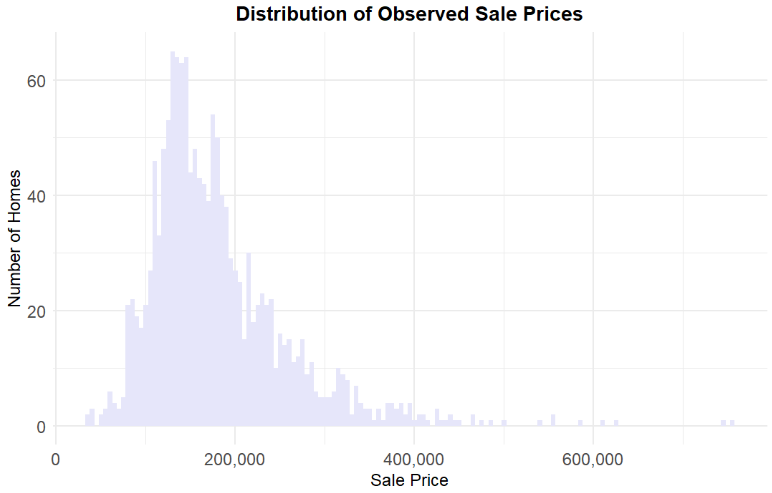
After each model was trained, the importance of each variable utilized by the model was measured. The resounding most important feature was "OverallQual", or the overall material and finish quality of the observed dwelling. The figure to the left shows the list of features with a greater than 1% importance to the neural net model.



Using the predictors above my neural network model was able to predict prices for the test set data.

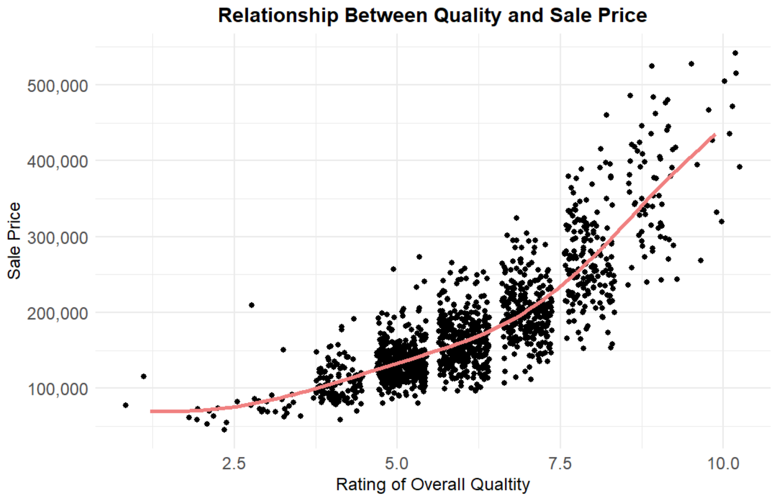


The model above presents visualization of the predicted sale prices of homes, utilizing the data obtained from our neural network model. The width of each bin represents $5,000; the bar's height shows the number of homes within a specific price range. The distribution was heavily skewed to the right – which makes sense for housing price data. The mean home price predicted was $181,179. Observed sale prices from the training data set are shown below.



The observed sale prices were skewed farther to the right, with a maximum value of $755,000 compared to the predicted price maximum of $554,145. The mean value of the observed sale prices was $180,921. Overall, the distribution and summary statistics of the predicted data - compared to the actual observed prices - gives me confidence in the explanatory variables used in my model.

To explore overall quality as the premier feature of my model, the following graph was designed to give a visual representation of the relationship. Overall quality is logically expected to positively influence the sale price. While there are some outliers, the general trend indicates that higher home quality typically commands a higher purchase price. Notably, homes rated 9 or above had a minimum sale price that was at least 50% greater than the average sale price of homes rated below 9. Similarly, the average sale price of homes with large basements (top 25% in square footage) is roughly 67% higher than the average sale price of homes without large basements. The factors previously mentioned are likely contributing to the skewed shape of the sale price distribution.



Conclusion

In summary, the sale price of homes in Ames Iowa is largely driven by space and quality of finish. There are several factors, like gabled roofs and cinderblock foundations, that negatively impact sale prices. For the most part, though, the features found to be important to sale price were positively correlated. To reinforce this observation, recently built or remodeled homes tend to command higher selling prices compared to older properties that could benefit from renovations.

Interestingly, when considered in isolation, some neighborhoods show a negative correlation to sale prices. That effect, however, is of little consequence compared to space and quality. A well-kept or renovated home in a less than desirable neighborhood still fetches a competitive sale price. While some neighborhoods did have a positive influence on price, these effects were generally minimal.

In the future, modeling may benefit from better preprocessing – increasing model prediction power. Box-Cox and Yeo Johnson preprocessing methods would be ideal but the effort to convert predictions into tangible results should be considered.