Springboard -- Data Science Career Program

Capstone Project #1 - Real Estate Price Predictor

Final Report

By Kevin Cole

May, 2020

1. Introduction

In real estate, finding properties that are undervalued and avoiding properties that are overvalued can help the long term prospects of the investment. A predictive model of housing prices would then be valuable to help investors identify valuable or mispriced properties and sellers to price their properties appropriately. This project seeks to create predictive models for sale price for homes in the Ames, Iowa housing market.

Initial proposal, milestone reports, and implementation of techniques can be found at https://github.com/ABitNutty/Capstone-1

2. Approach

2.1 Data Acquisition and Wrangling

The data: The data set is a premade .CSV file found on the Kaggle website. The dataset contains 1460 data points, each being the records for a property sold in Ames, Iowa. There are 79 variables available for analysis, along with a column for the target variable of sale price.

Data cleaning/wrangling:

- Variables were investigated and found to be appropriately labeled, and the index was set to “ID”.
- Variables were categorized as either categorical or non-categorial.
- Dataframe was organized by tidy data standards.
- Categorical variables were investigated through histograms and frequency counts.
- Numerical variables were investigated through basic summary statistics.
- Missing values were imputed on a case by case basis.
- Dummy variables were created in replacement of categorical variables.

2.2 Storytelling and Inferential Statistics

Housing prices over time: Time is one of the factors that can drastically influence the price of a home. For this market, it would be useful to understand the general trend of the housing market and how much you can expect property values to rise year over year.

Figure 1: Sale price over time

In figure 1, we can see a slight trend that does suggest the prices are increasing over time for new constructions. To quantify rates, we look at the regression line for the entire dataset, as well as that of properties since 1990. We can see in Figure 2 the increasing trend in the data, measured at $1375/yr overall and $2526/yr since 1990. This suggests that the new construction housing market in the area is not only increasing but has accelerated in the last decades.

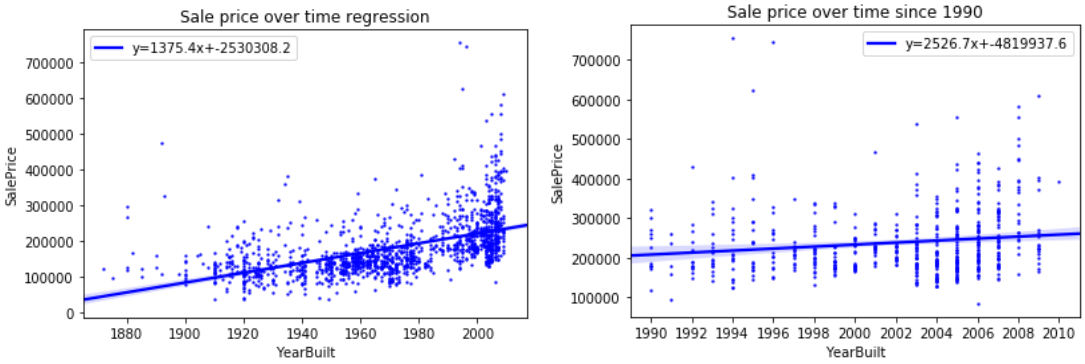

Figure 2: Linear regression of sale price over time.

Strong predictors of sale price: Understanding the strongest predictors of sale price can influence the investment/sale of investment decisions or even new construction decisions. A

frequently provided metric in housing ads is garage spots. In figure 3 we look at a comparison of sale price to garage spots. While not exactly an independent predictor of sale price, we can see that for a house to achieve the top value of properties in the area it is almost mandatory to have 2-3 garage bays. This would be one of the high value items a new construction in particular could pay attention to in order to achieve a higher sale price.

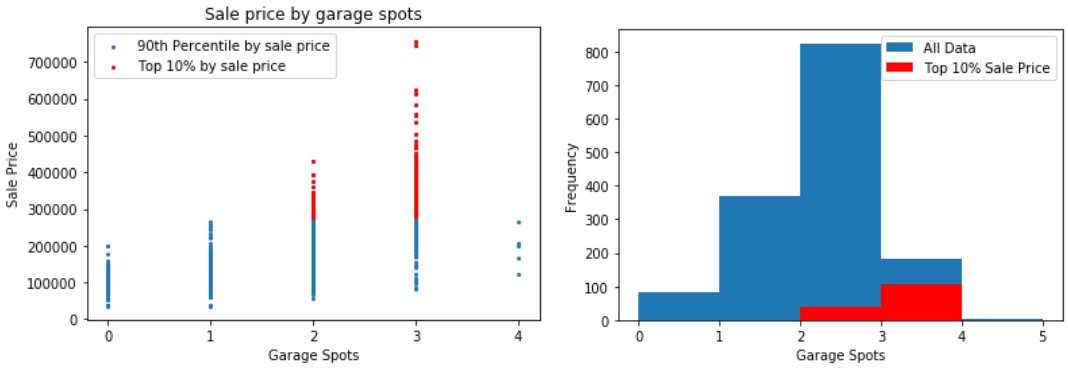

Figure 3: Sale price by garage spot

In looking for variables with a high correlation with sale price, a heatmap was used where a correlation coefficient of 0.6 or higher was deemed a strong predictor.

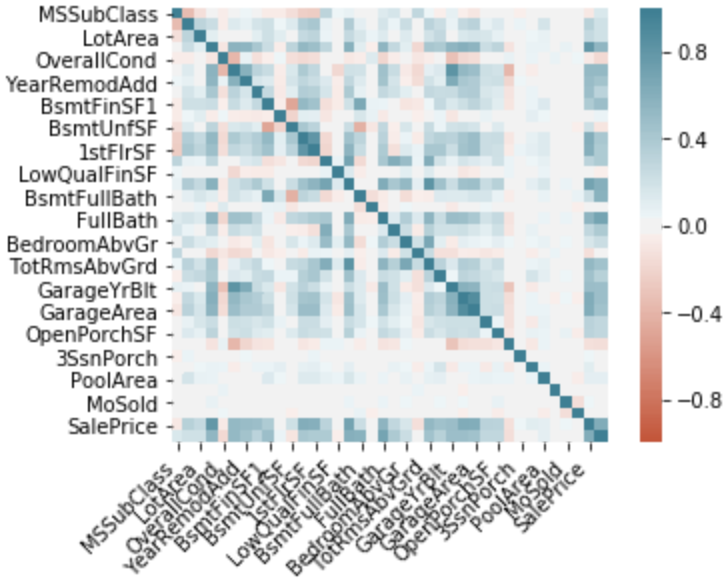

Figure 4: Pearson Correlation Coefficients heatmap

In the investigation of correlation coefficients we find the top three variables are overall quality, square footage of living area, and number of cars in the garage, followed by garage area, and square footage of basement and first floor.

The garage qualities are showing some strong influence on sale price. We can hypothesize that the mean sale price for attached garages is equal to that of detached garages. A histogram of occurrences (Figure 5) for our sample we notice that in this particular sample there is a difference between the means of the subpopulations.

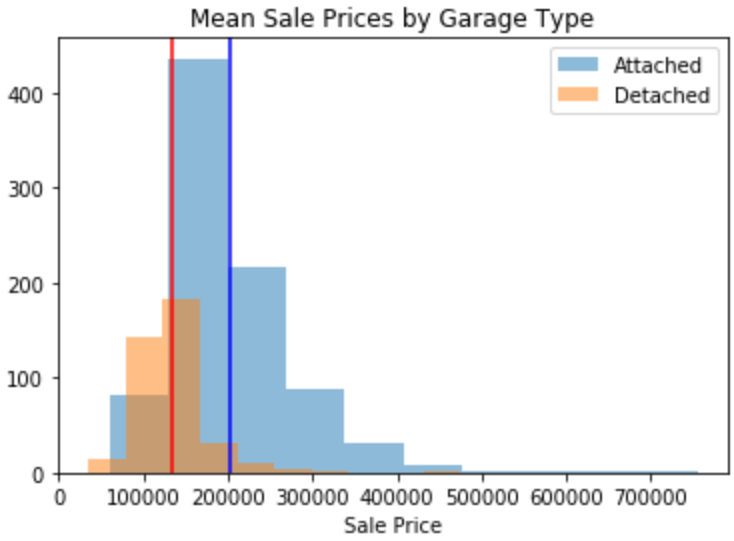

Figure 5: Attached vs. Detached Garages

By creating bootstrap replicates of the difference of means in each population, zero events were found to have a difference in means at least as extreme as what occurred in our sample. We can then conclude with high confidence that the difference observed in sale price for attached vs. detached garages is statistically significant.

2.3 Baseline Modeling

The baseline model was measured against three metrics. R^2, the traditional correlation coefficient for a linear model, RMSE (Root Mean Squared Error), to give a sense of spread of the errors, and MAPE (Mean Actual Percent Error), a measure of accuracy as a percentage.

Several models were tested, starting with a standard linear regression, followed by Ridge and Lasso to examine models for overfitting. Ridge regression provided the least amount of overfitting to the train test split (25% test data).

Hyperparameter tuning of the Ridge model produced the best results with an alpha (regularization parameter) of 10. Outliers were investigated and there was no practical reason to remove them from the baseline model.

Figure 6: Baseline model predictions

2.4 Extended Modeling

A Random Forest Regressor model was trained using the same train/test split as with the linear regressor. Tuning of the hyperparameters led to a model with max tree depth of 10 with 100 trees in the forest and 50 maximum features. When outliers to the model were removed, the same tuning results were achieved. Outliers were left in the final model for results measurement.

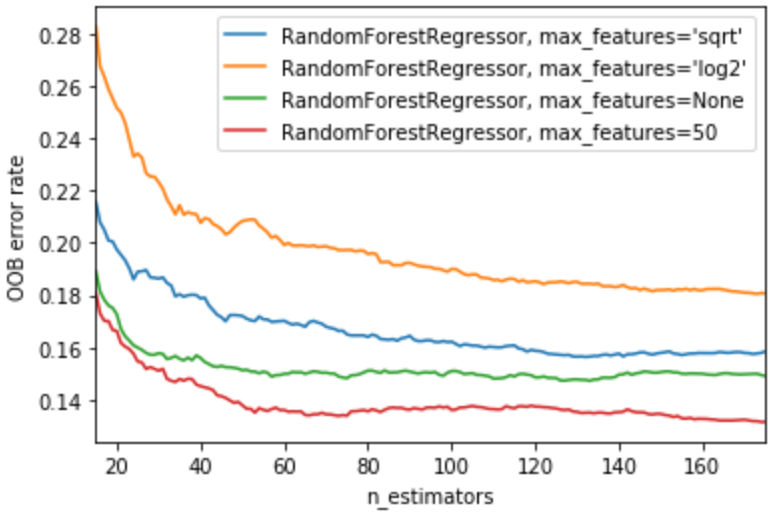

Figure 7: Tuning of features

3. Findings

After each model was run, performance metrics were recorded. In both instances of outlier removal, the model did not improve and those models will not be the final model. Figure 8 shows the performance metrics for the tests sets of the various models run as well as the prediction plot for the model chosen. As mentioned in the baseline section, the Ridge regressor provided the least overfitting, and the Random Forest Regressor provided the best overall results as seen in the 'Forest_Test_Tuned' row. Worst underestimates and worst overestimates are the worst case scenarios the model can be expected to produce given a 95% confidence level.

Model	R_Squared	RMSE	MAPE (%)	Worst Underestimate	Worst Overestimate
Linear_Test	0.877025	27774.277801	11.107240	54799.1	55049.2
Ridge_Test	0.894620	25710.742519	9.869781	40291	47665.9
Lasso_Test	0.887081	26614.446974	50.809840	219702	219698
Ridge_Test_a=10	0.659501	46216.022922	18.516758	90024.5	75831.5
Ridge_Test_No_Outliers	0.649496	40840.354214	16.339041	83885.2	72174.1
Forest_Test_Initial	0.890738	26180.002763	10.020665	51054.8	49975
Forest_Test_Tuned	0.897620	25342.031965	9.991466	49009.8	48301.1
Forest_Test_N_O	0.874338	28193.940135	11.251358	54936.3	51446.4

Figure 8: Performance Metrics

Figure 9: Final Model - Random Forest (test set)

This model can be used by an investor in the area looking to predict the sale price of a home. This prediction, and relative error, can be used to find undervalued properties for purchase and investment or to prevent significantly overpaying for a property.

4. **Conclusions and Future Work**

Feature importance: Feature importance to the model can drive many business decisions. The features most influential to the random forest model are shown in Figure 10.

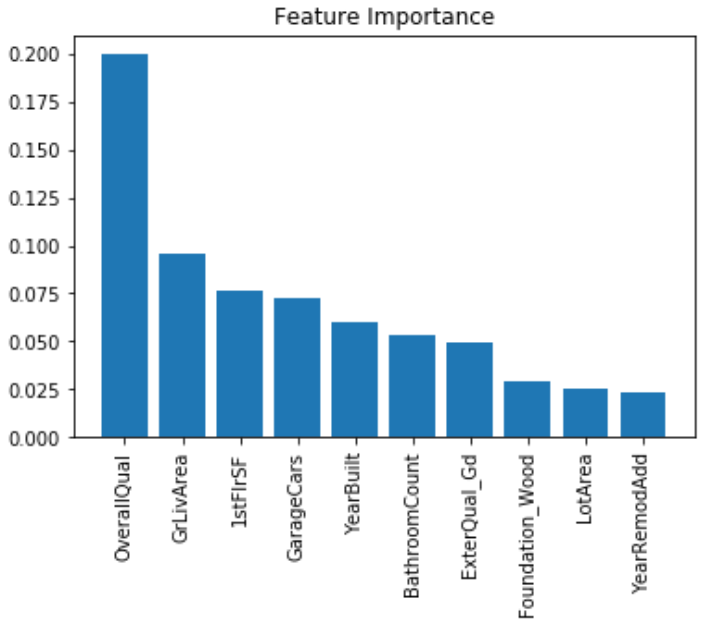

Figure 10: Feature Importance

By seeing just how important the overall quality rating is to the sale price, almost twice that of any other variable, we can conclude that a person's first overall impressions are vitally important to their valuation of the property. This is particularly important if that person is a buyer. This is valuable to realtors and house sellers as they can make decisions on decor and presentation that can artificially inflate its perceived value before an offer is made.

Garage space and living areas are also highly important to the sale price, which is valuable for builders and developers as they design new constructions. Maximization of those features could lead to a higher sale price and a higher return on investment.

Future work could include:

- Different subsets of the original feature set - given the detections of important features discovered in these models, tuning of the feature set could be strengthened.
- Other models such as Neural Nets could be built.
- Expansion of the model to cover nearby markets or different markets all together.

5. Recommendations for the Clients

- Having understood through the models built that housing sales are very much dependent on first impressions, any business actions that can increase a buyer's initial perception of the property will directly lead to a higher sale price. This could include upgrades like paint or flooring prior to listing, or staging efforts at open houses.
- When identifying undervalued properties, the model should be producing a predicted price higher than listed. Due diligence is still required in these instances as there can always be more to the story for an individual property.
- Identifying properties for investment should include not just properties where the model has produced a predicted price higher than listing, but those who score poorly in the high feature importance categories. Identifying categories that are easy to upgrade can lead to fast turn-arounds and remodels.
- When designing new constructions, the coefficients of the linear regression model on the z-scored training set would give you an idea on how the price would grow as a function of features such as number of garages, bathrooms, or square footage. Planning these features into the design can directly lead to a more valuable property.
- Continue to feed and refine the model if it is indeed to be used over time. The model was built on a relatively small data set and more data could prove valuable. The model should not be used in other markets until trained on data from that market as market conditions can vary greatly by region.

6. Consulted Resources

- Kaggle - initial dataset - https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- Data Camp -- Python skills and models influenced by data camp exercises. List of courses used can be seen at https://www.datacamp.com/profile/kevinccole
- CS109 lectures provided via Springboard curriculum.
- Documentation for imported packages.
- Stack Overflow for individual coding questions.
- Consultations with Data Science professional AJ Sanchez.