Recommendations for Remodeling

By : Ben McPeek

Problem Statement

A construction company reached out to Ames Assessor's Office to get insight on what kind of remodeling they could recommend that would best increase the sale price of a house. We will be creating a Linear Regression model that will show the kinds of remodeling that best increase sale price and the ones that show the least impact on sale price. The features in our model will incorporate standard strong linear relationships with sale price such as overal_quality, overall-cond, and year_built as well as columns that are involved with remodeling. We will evaluate the success of each feature by interpreting their coefficients.

Data Description

This data set contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. The original data (obtained directly from the Ames Assessor's Office) is used for tax assessment purposes but lends itself directly to the prediction of home selling prices. The type of information contained in the data is similar to what a typical home buyer would want to know before making a purchase and students should find most variables straightforward and understandable.

DATASET SHAPE:

Train = (2051, 81)

Testing = (878, 80)

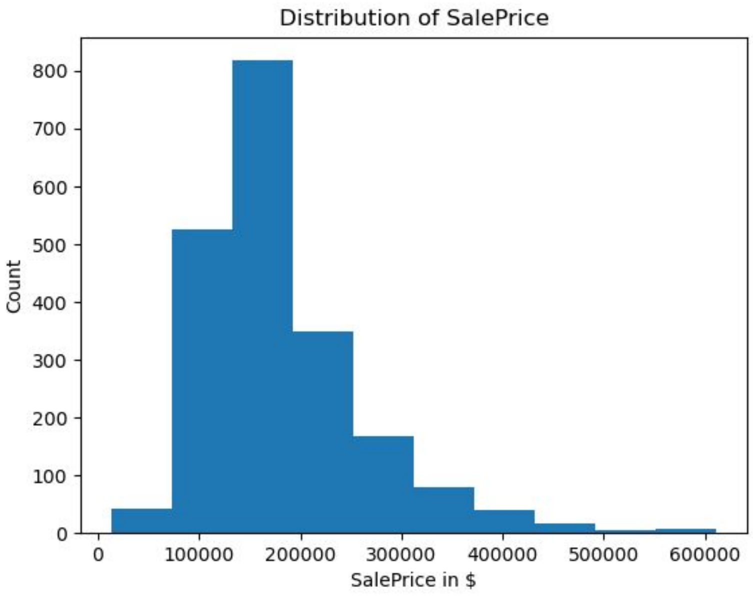
Target Variable = 'Sale Price'

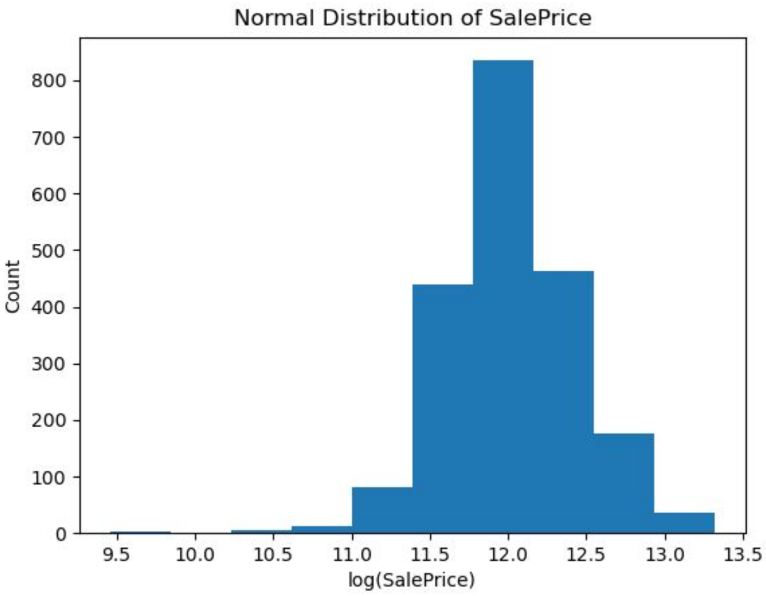
Sample Description of Features

Feature	Type	Dataset	Description
lot_area	*float*	Train Dataset	Lot size in square feet.
overall_qual	*integer*	Train Dataset	Rates the overall material and finish of the house.
year_built	*integer*	Train Dataset	Original construction date.
year_remod/add	*integer*	Train Dataset	Remodel date (same as construction date if no remodeling or additions).
full_bath	*integer*	Train Dataset	Full bathrooms above grade (ground).
1st_flr_sf	*float*	Train Dataset	First Floor square feet.
gr_liv_area	*float*	Train Dataset	Above grade (ground) living area square feet.
house_style	*integer*	Train Dataset	Style of dwelling.

Distributions

Sale price has a right skew distribution. By using log method we can create a normal distribution



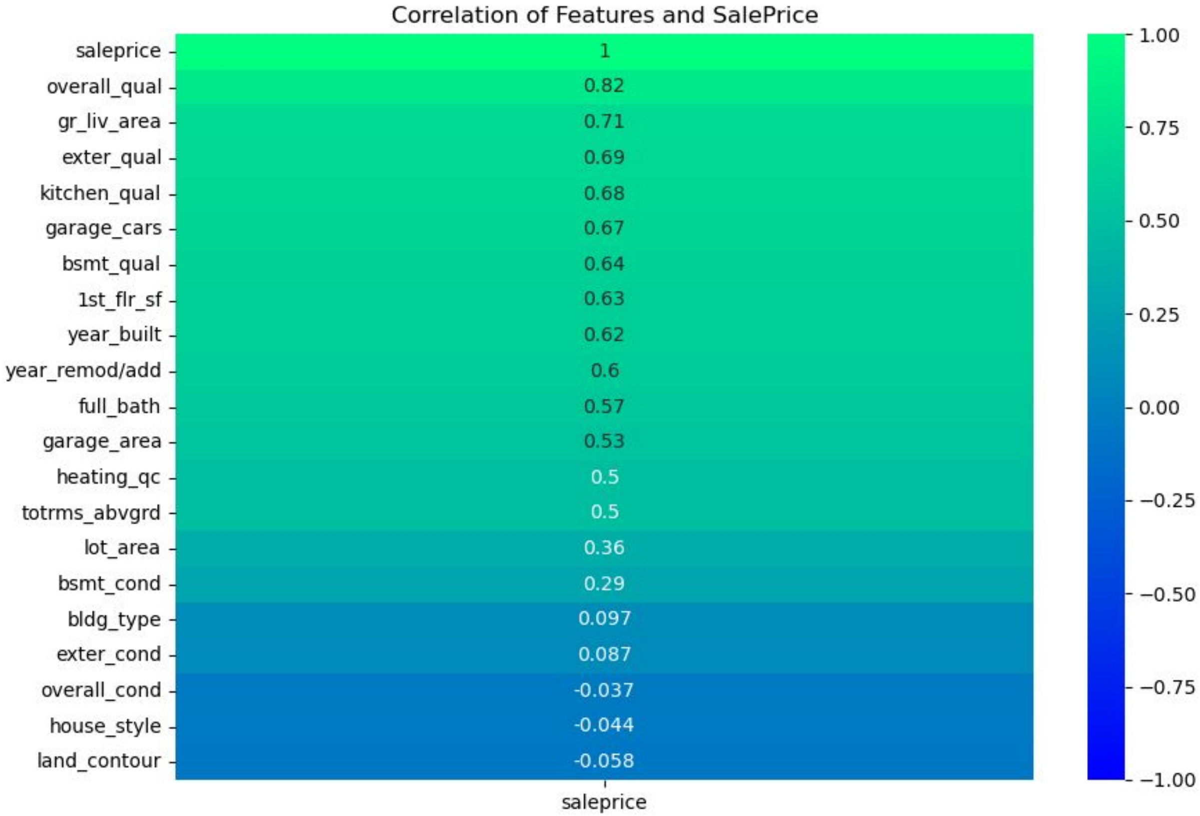


Searching through data dictionary

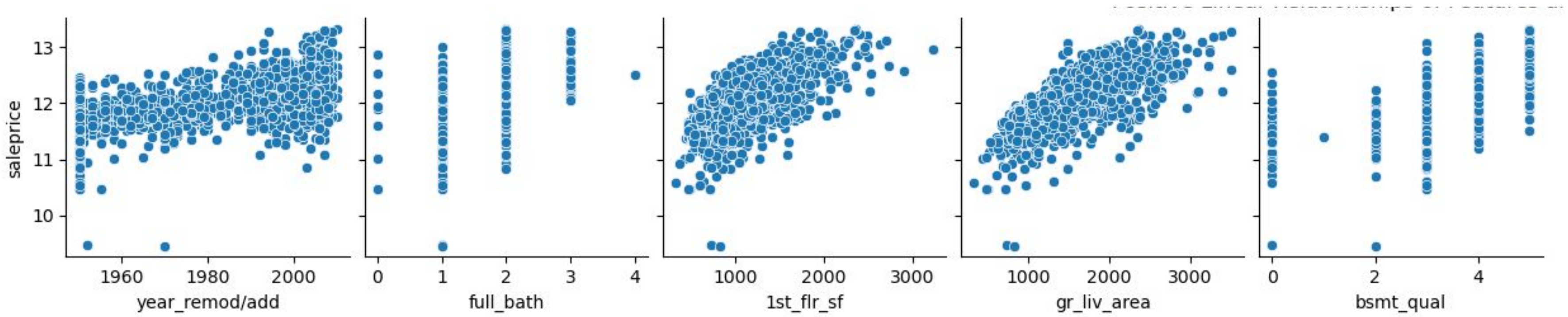
features = ['bsmt_qual', 'bsmt_cond', 'heating_qc', 'kitchen_qual', 'garage_area', 'totrms_abvgrd', 'full_bath', '1st_flr_sf', 'gr_liv_area', 'garage_cars', 'lot_shape', 'land_contour', 'bldg_type', 'house_style', 'foundation']

Looking through the data dictionary, I found features that related to my problem and correlated with sale price. Columns such as 'year_remod/add', 'bsmt_qual', 'bsmt_cond', 'kitchen_qual', 'garage_area', 'full_bath', '1st_flr_sf', 'land_contour', 'bldg_type', 'house_style' are features that have correlations with sale price that are very insightful for learning how and what kind of remodeling best increases the sale price of houses.

Looking at our correlations we can see some of the features I selected holding strong relationships with our target variable.



Scatter plots relating Sale Price to some of our features confirms our correlations. We can see strong positive linear relationships in these scatter plots.



Linear Regression, LASSO, and Ridge Models

- I ran my features through a linear regression model and used StandardScalar to balance the values of my data (features had a variety of value types from continuous to discrete). Using train_test_split I got a cross_val_score of 0.884 which shows my features fit well to my model.
- Training R2: 0.895, Test R2: 0.875 shows our features fit well with our model and because our test and train scores are close enough, it shows our model will pass through new data fairly accurately.
- Our lr model showed us that our features had a RMSE of $196097.015 which means our predictions deviated higher to the observed value than we wanted.
- I incorporated more features into a lasso and ridge model to see if I could increase the model performance higher.
- The ridge model ended up performing the best with R2 scores of Train: 0.913, Test: 0.877.

Conclusion

Features with the highest coefficient

gr_liv_area | 0.109882 | $109.88

- For every 1 square foot increase of the above ground floors there is a roughly $109.88 increase of sale price of the home. (Given all other things held constant)

1st_flr_sf | 0.068947 | $68.94

- For every 1 square foot increase of the ground floor there is a roughly $68.94 increase of sale price of the home. (Given all other things held constant)

bsmt_qual | 0.027507 | $27.51

- If the basement quality raises one grade level we can expect a $27.51 raise in sale price of the home. (Given all other things held constant)

kitchen_qual | 0.022340 | $22.34

- If the kitchen quality raises one grade level we can expect a $22.34 raise in sale price of the home. (Given all other things held constant)

Recommendations

- These four features were shown to positively affect the sale price of homes the best. Our model recommends that the construction company should advise their clients in two ways. If their client is looking for heavy remodeling, expanding the square footage of the ground or upper living areas would best increase the sale price of their home. If a client is only interested in minor remodeling, upgrading the quality of their kitchen or basement would be the most effective way.
- Features such as 'lot_shape', 'garage_area', and 'house_style' showed very low effect on the sale price of houses. It is recommended that the construction company advise their clients interested in increasing the value of their home to not invest in the shape or style of their house or the size of their garage.
- These models could be applied to other cities but the features would need to be altered depending on what remodeling is common in those areas.