Ames Housing Sale Price Prediction

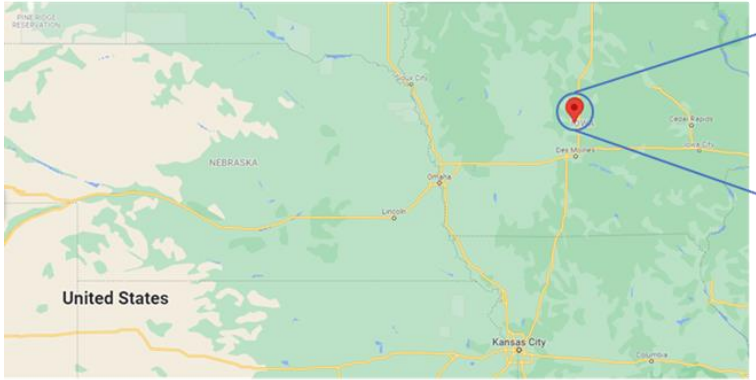
Team 5: Bryan Soh
Mitchelle Chua
Vincent Chua
Zavier Soon

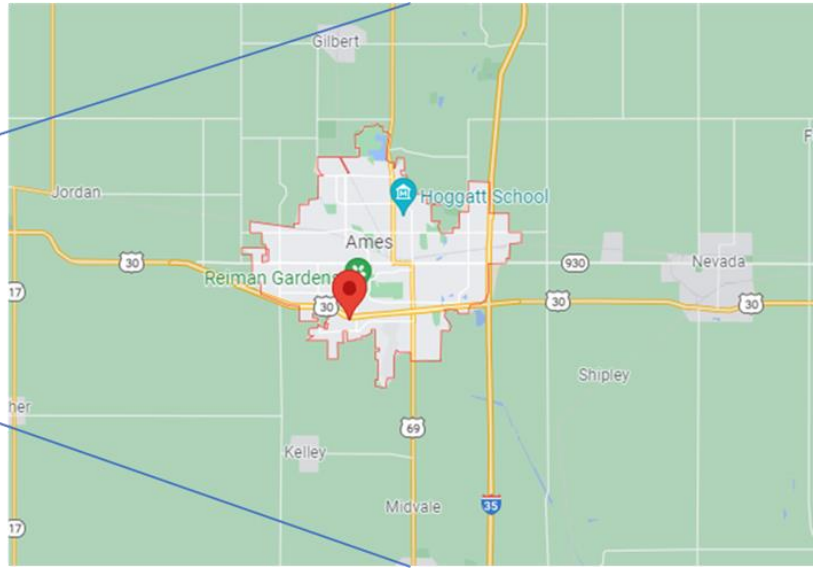
Ames, Iowa

Official Website: https://www.cityofames.org

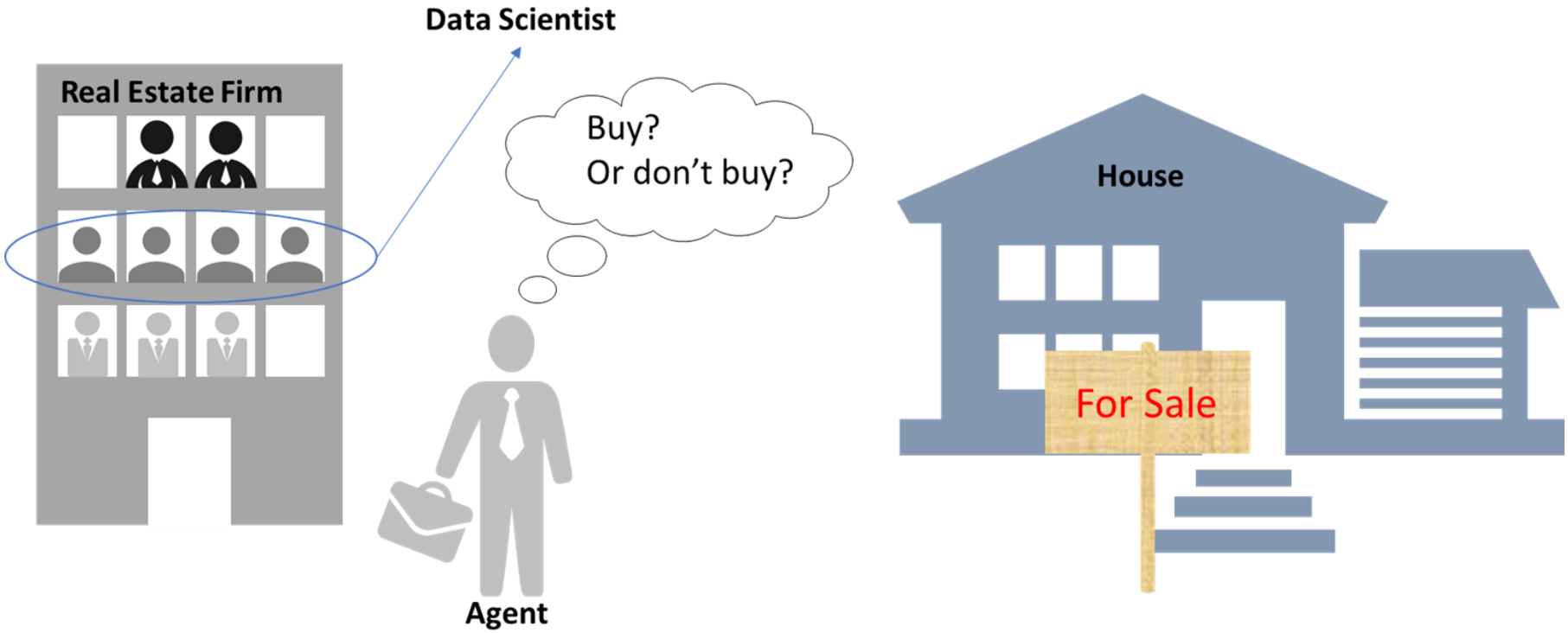
Ames is a city in Story County, Iowa, United States, located approximately 30 miles north of Des Moines in Central Iowa. It is best known as the home of Iowa State University, with leading agriculture, design, engineering and veterinary medicine colleges.

From: Google Map





Background


Data Scientist
Real Estate Firm
Buy?
Or don't buy?
House
For Sale
Agent

Problem Statement

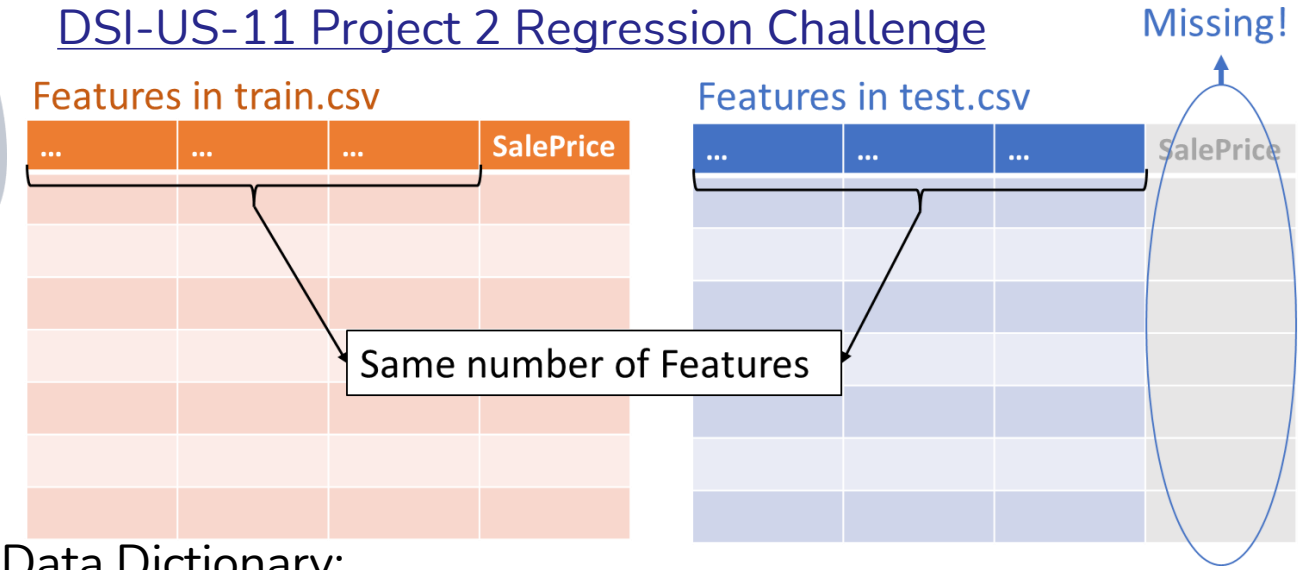
To explore and analyse the dataset to develop a model that predicts the housing sale price in Ames, Iowa.

Using the model built, identify

- 3 features that will increase the sale price
- 3 features that will lead to a decrease in the sale price
- 3 features recommended for renovation

Link to data (train.csv and test.csv)

[DSI-US-11 Project 2 Regression Challenge]



Data Dictionary:

https://web.archive.org/web/20201203235151/http://jse.amstat.org/v19n3/Decock/DataDocumentation.txt

Feature and Data imputation

Features dropped:

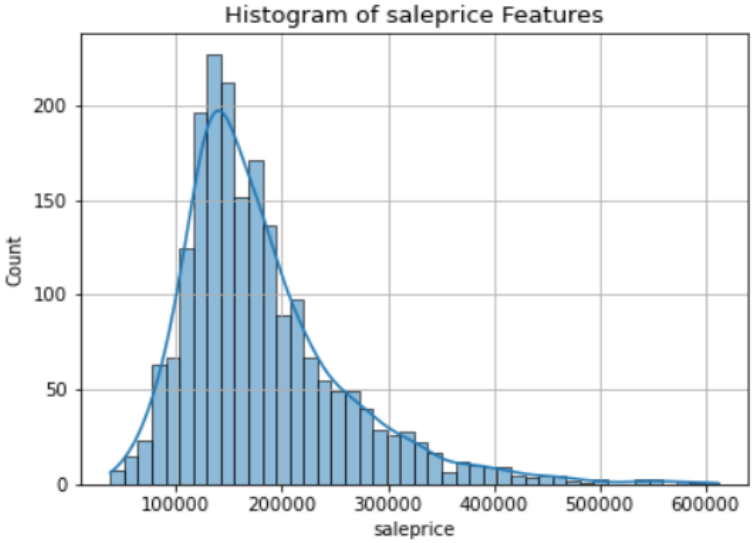
1. Lot Frontage (more than 5% of the data are missing)

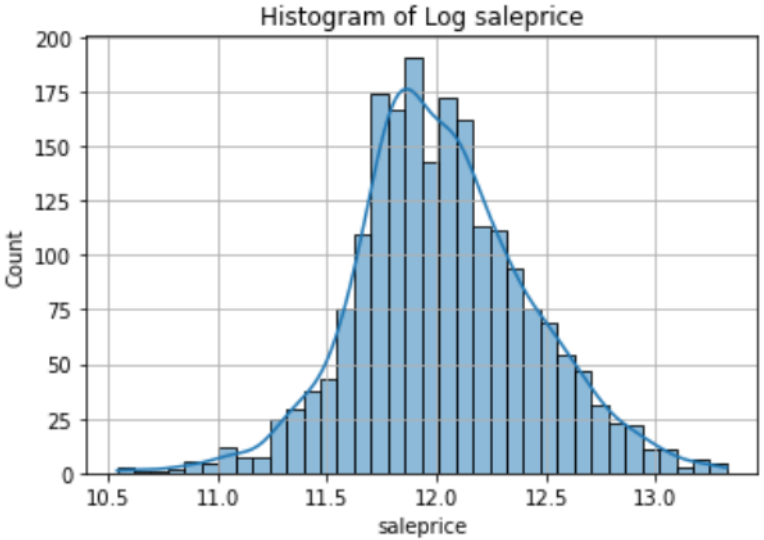
Data dropped in train.csv

1. Data with lots of missing values
2. Data with extreme values (outliers)

Feature Engineering

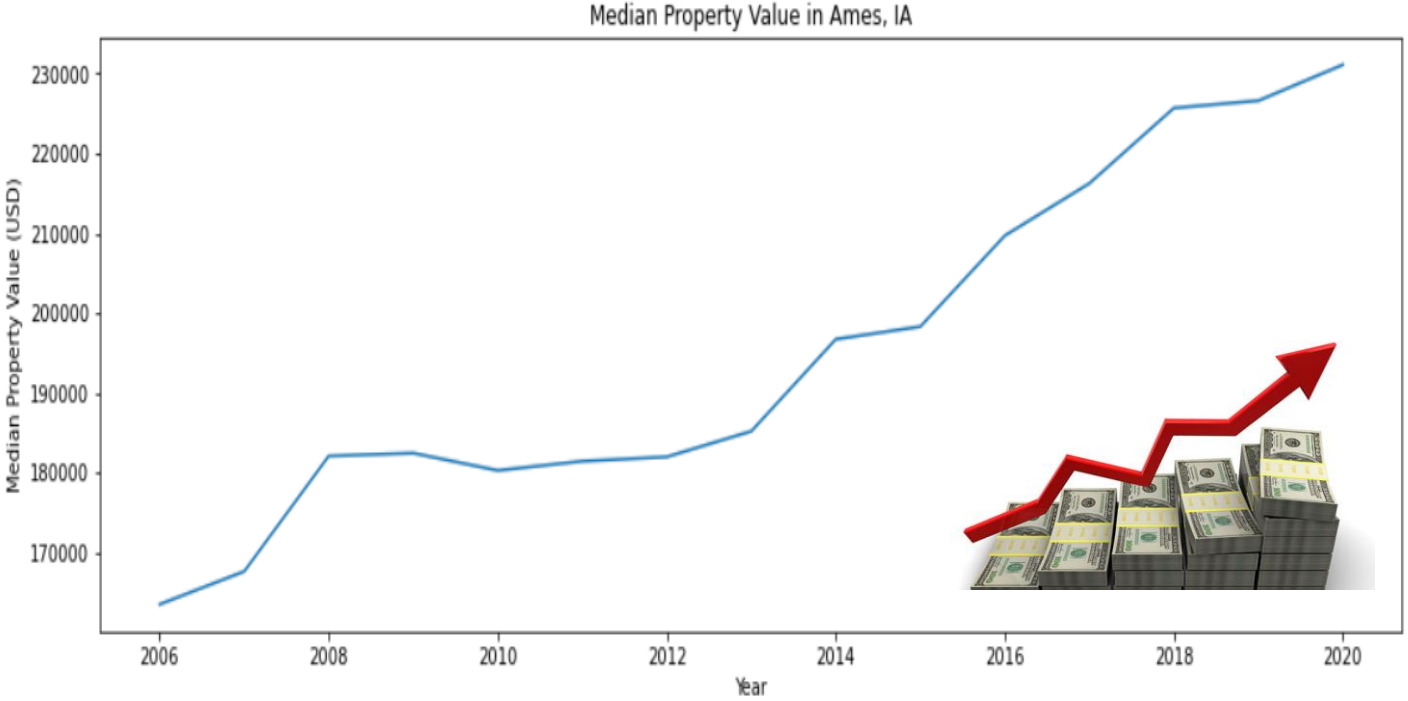
1. Saleprice Log Transformed





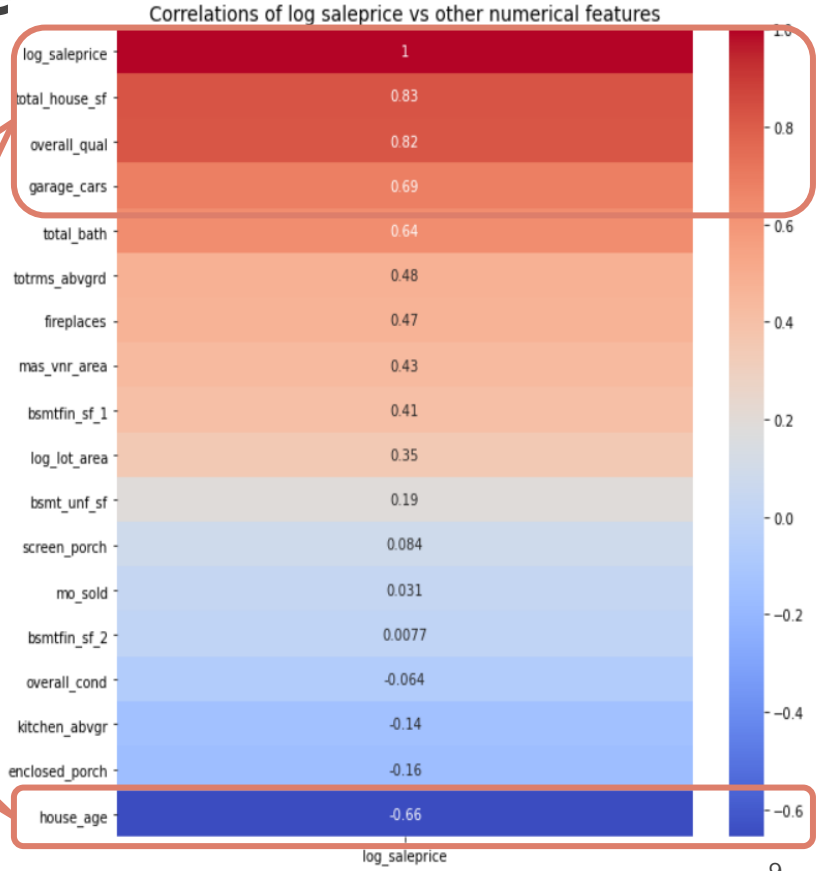
1. Three New Features:
   - Total square feet of a House
   - Quality Score
   - House Age

Median Property Value in Ames, IA (2006-2020)

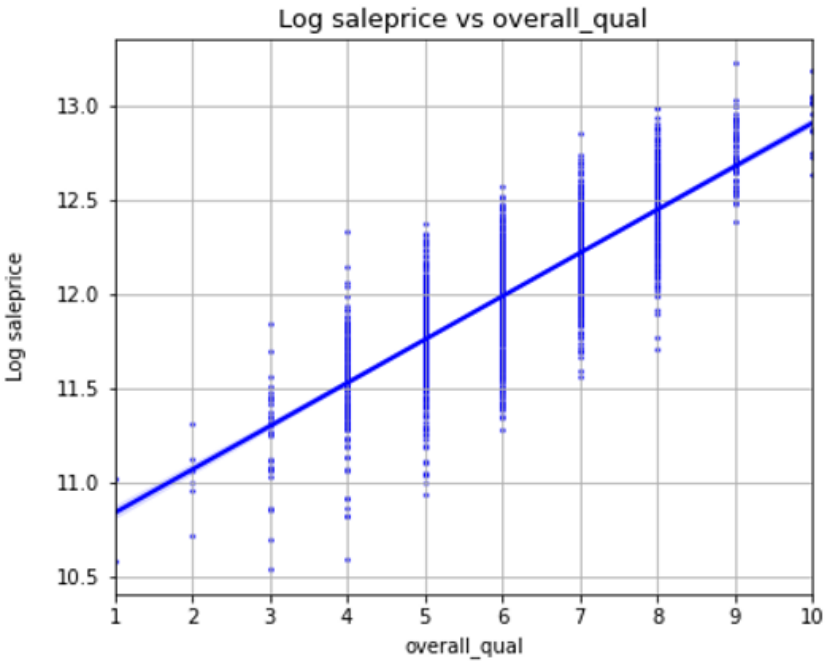


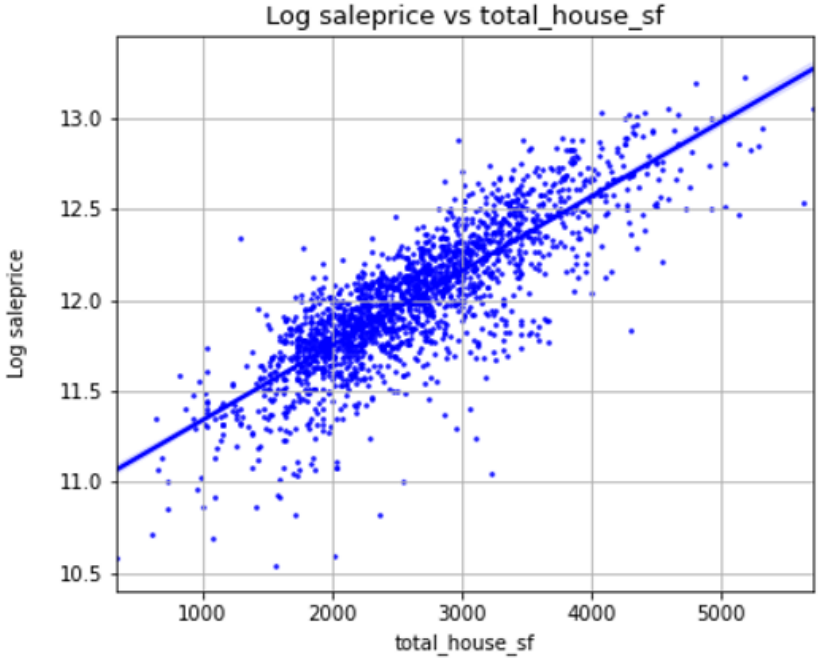
References:
1. https://walletinvestor.com/real-estate-forecast/ia/story/ames-housing-market
2. https://data.census.gov/cedsci/table?q=DP04&tid=ACSDP5Y2019.DP04

Correlations of Sale Price vs Other Features

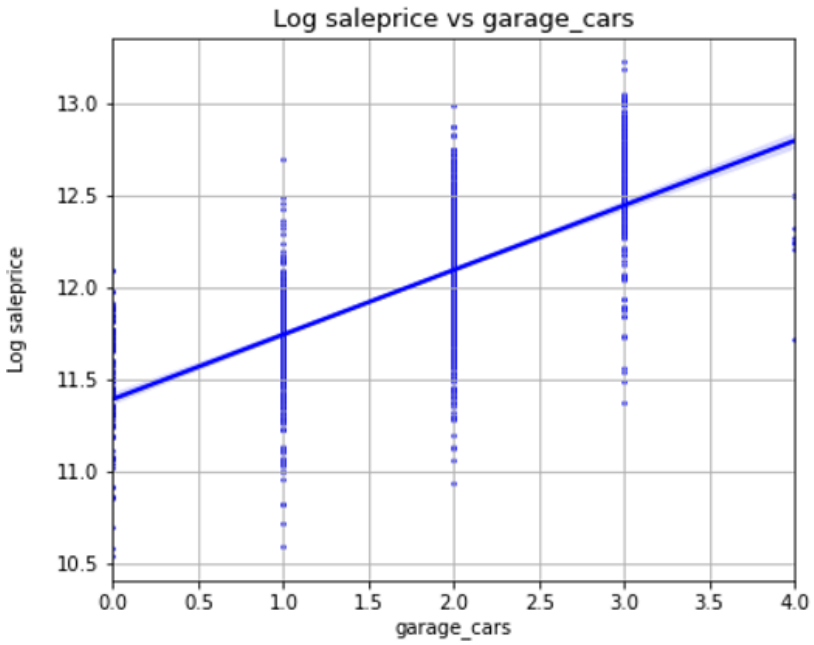


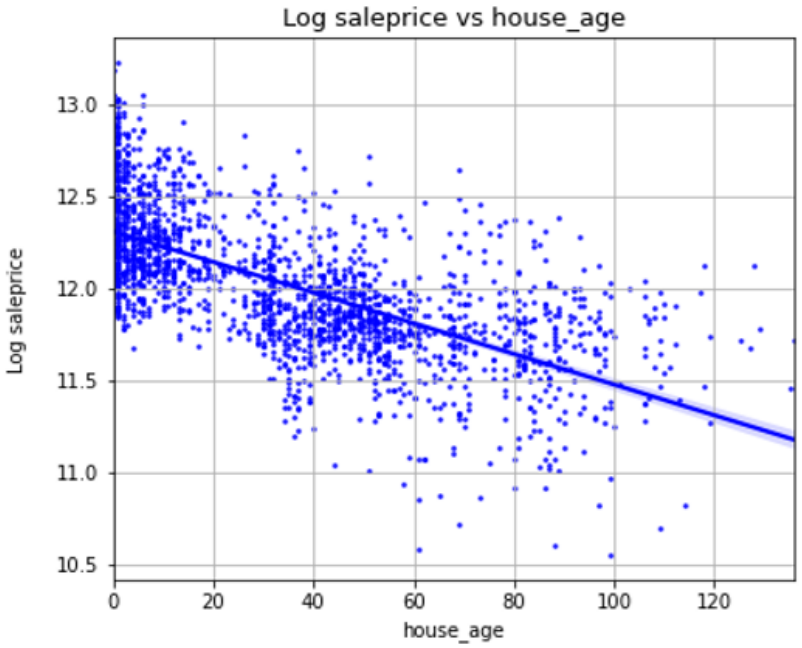
Top 4 Features Highly Correlated Sale Price



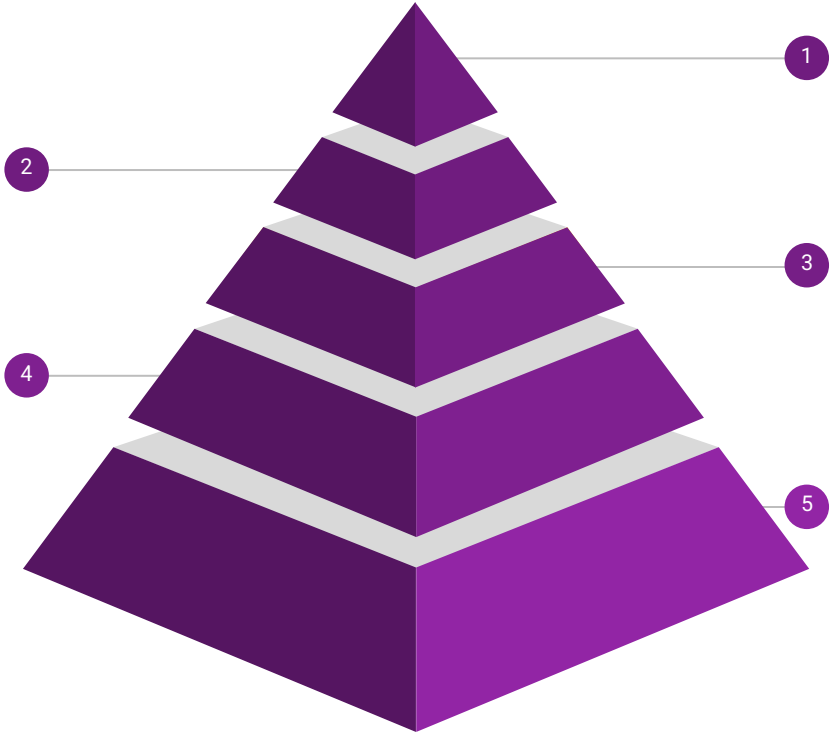


Top 4 Features Highly Correlated Sale Price





Our Modelling Process

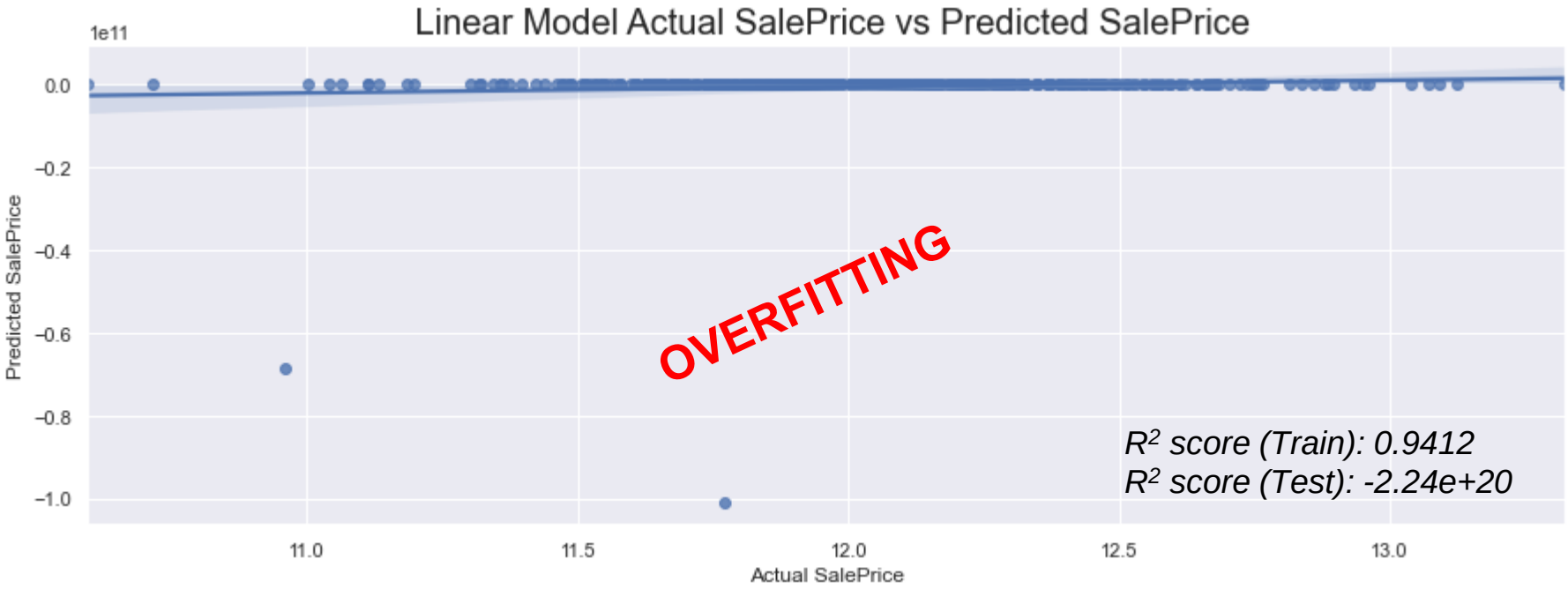


1. **Final Model & Prediction**
   Use our selected model for prediction
2. **Model Evaluation & Selection**
   Use evaluation metrics R^2 Score and Root Mean Square Error (RMSE)
3. **Regularization Model**
   Ridge Regression & Lasso Regression
4. **Baseline Model**
   Linear Regression
5. **Model Preparation**
   One Hot Encoding, Scaling, Train-Test-Split

Model Preparation

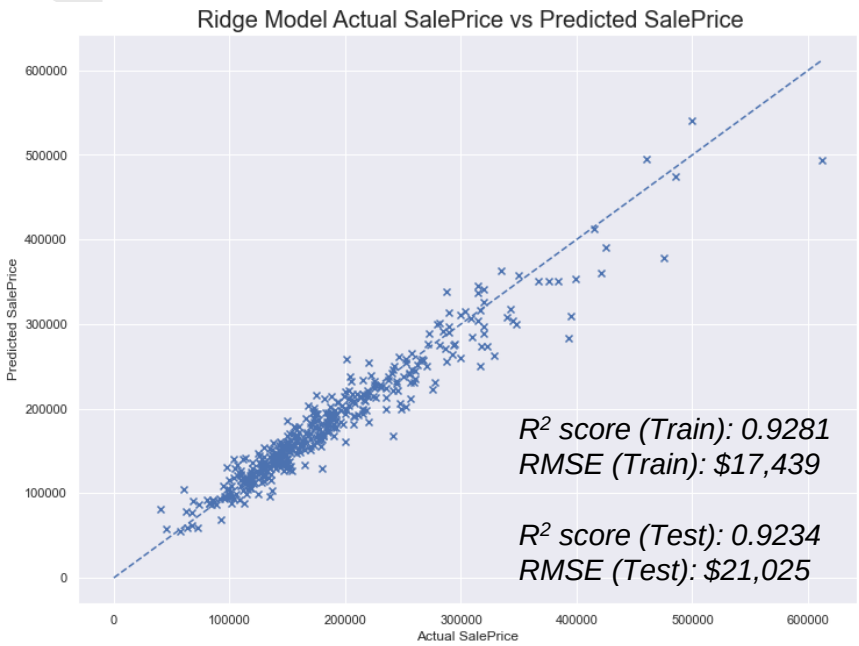
- One-Hot Encoding our categorical variables
- Moving from 77 features to 235 features
- Create more noises and increase risk of overfitting
- Scaling our features
- Train-Test Split, 80% of train datasets

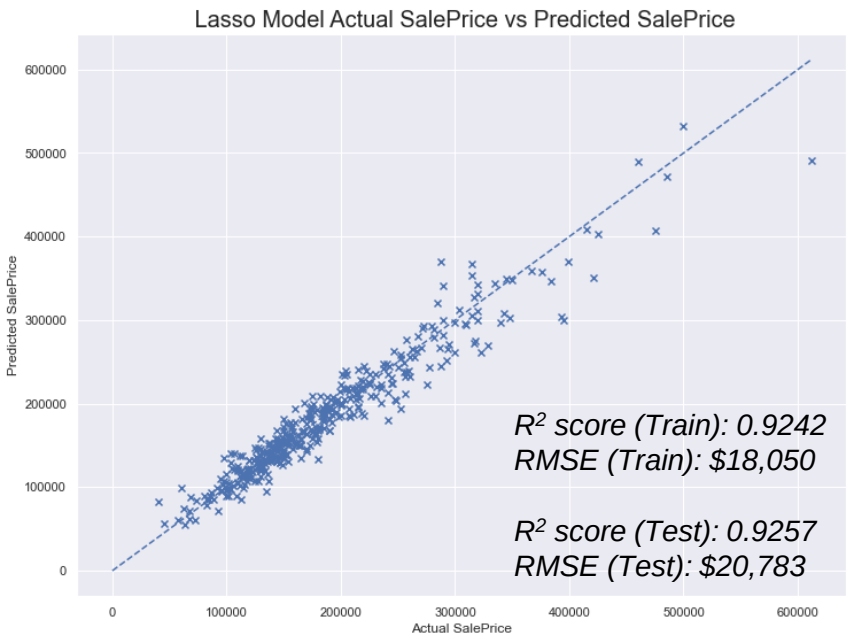
Baseline Model: Linear Regression



★ Coefficient of Determination **(R^2)** - This is to explain the accuracy of our model fits the observed data. The higher the r-squared indicated a better fit for the model

Regularization Model: Ridge vs Lasso Regression



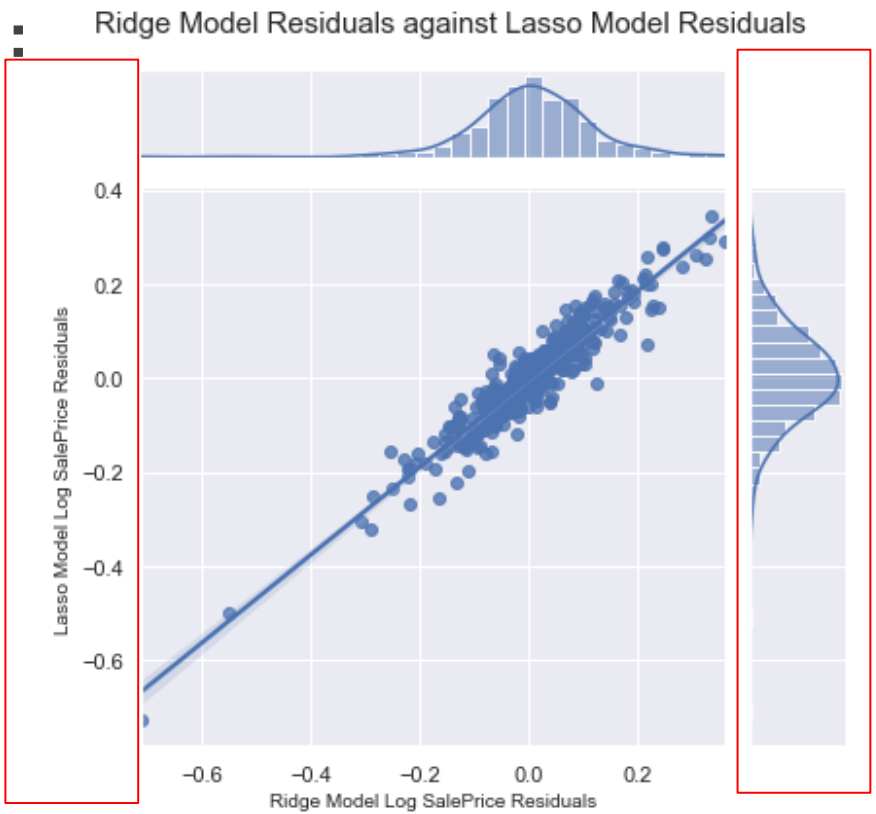


- ★ Coefficient of Determination **(R^2)** - This is to explain the accuracy of our model fits the observed data. The higher the r-squared indicated a better fit for the model
- ★ Root Mean Squared Error **(RMSE) -** the square root of the variance of the residuals. Lower values of RMSE indicate better fit.


RIDGE?
LASSO?

Evaluation of Metrics:

	Ridge Regression	Lasso Regression
Cross Validation Score	90.25%	90.15%
R^2 Score (Train)	92.81%	92.42%
R^2 Score (Test)	92.35%	92.57% ↑ 0.22%
RMSE (Train)	$ 17,439	$ 18,050
RMSE (Test)	$ 21,025	$ 20,783 ↓ 1.15%



★ Residuals - Difference between actual sale price and

Making sense of the Lasso Model Results

92.5%

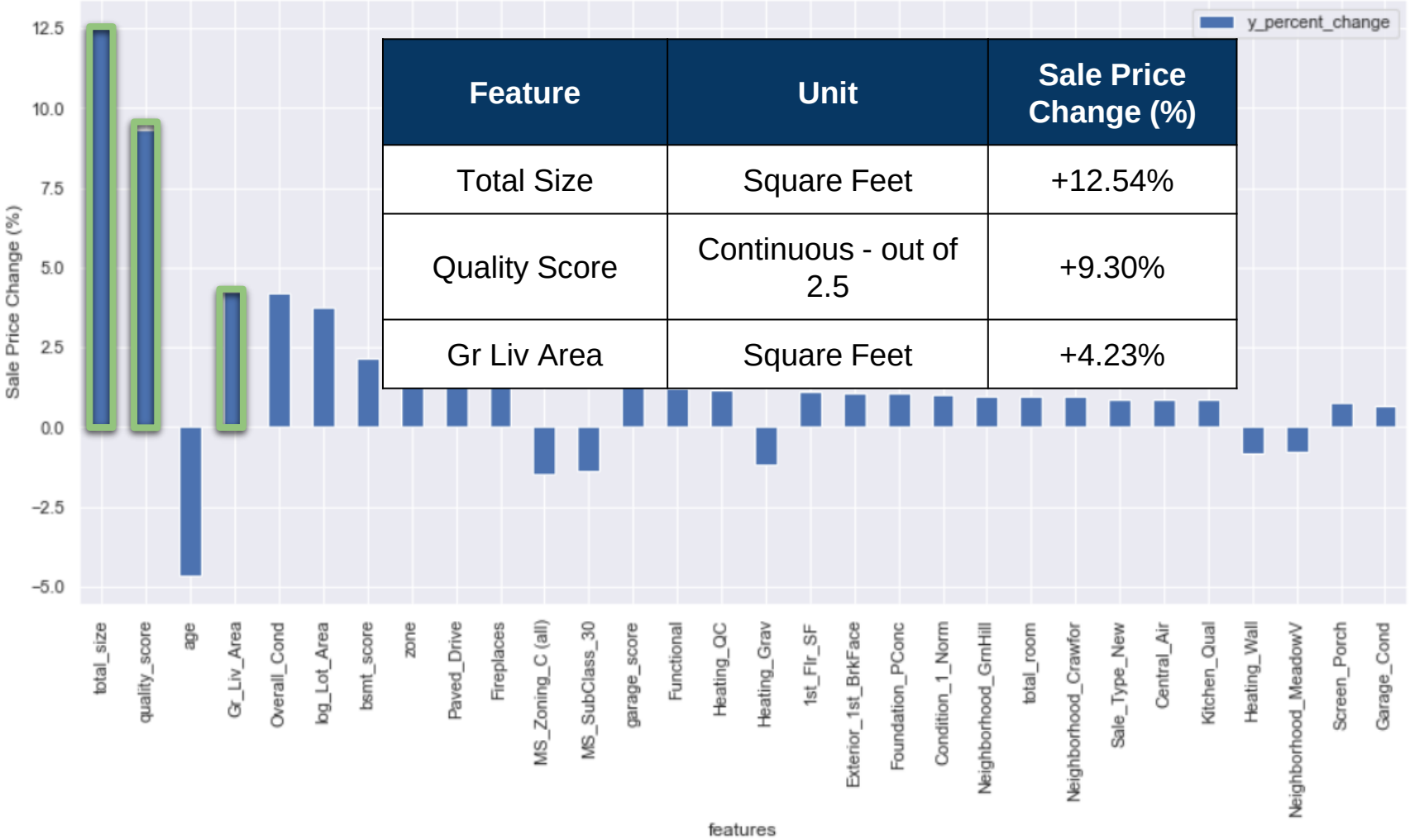
Variability or fluctuation in the test data sale price can be explained by the predictor features.

Making sense of the Lasso Model Results

USD 20,000

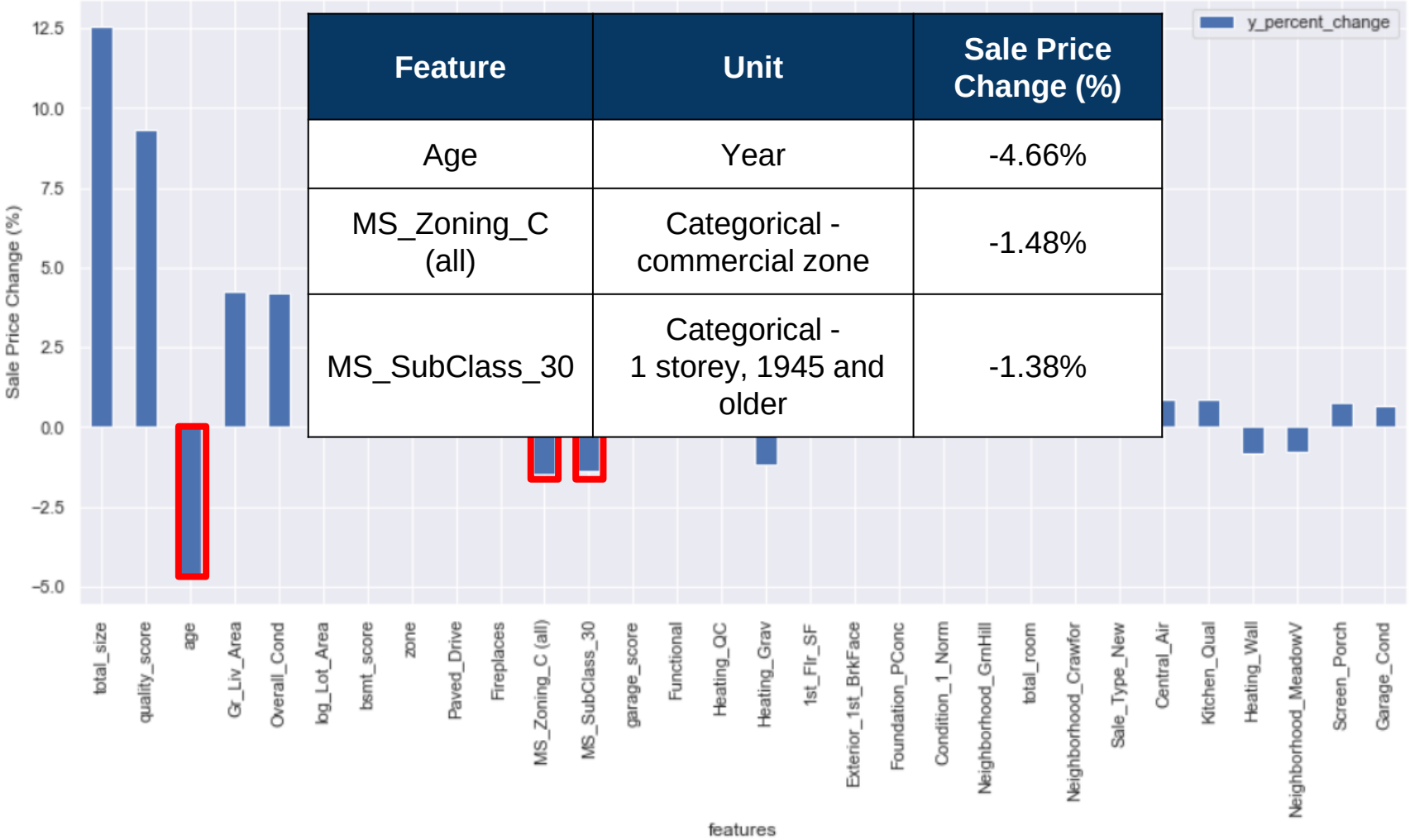
Predict the sale price within +/- USD 20,000

Top 30 Features on Housing Price Change in %



Feature	Unit	Sale Price Change (%)
Total Size	Square Feet	+12.54%
Quality Score	Continuous - out of 2.5	+9.30%
Gr Liv Area	Square Feet	+4.23%

Top 30 Features on Housing Price Change in %



Feature	Unit	Sale Price Change (%)
Age	Year	-4.66%
MS_Zoning_C (all)	Categorical - commercial zone	-1.48%
MS_SubClass_30	Categorical - 1 storey, 1945 and older	-1.38%

Top 3 Impactful Features for Renovation

Features	Breakdown	Levels	Impact per increase in level	Action
Overall Cond	Very Poor to Very Excellent	10	+4.21%	Repair and painting works
Paved Driveway	Dirt/Gravel to Fully Paved	3	+1.83%	Pave the full driveway
Home Functionality	Salvage Only to Typical Functionality	8	+1.22%	Repair damages

Conclusion & Recommendations

- High impact features are consistent with common sense, which lends credibility to the model
    - High quality, in good condition, newly built, large in size

- In terms of whether to renovate the house, the cost should be lower than the predicted increase in Sale Price

- The underlying guiding principle will always be to buy a house with a selling price lower than its true value