Predicting ~~for~~ Sale Price

Mandy McClintock
Machine Learning Project
NYC Data Science Academy
Feb 2021

The Challenge

For the Ames housing dataset, predict Sale Price.

Who **BENEFITS** from this task?

Sellers	It's important to know the value of your house. Rule #1: Do not overprice your home.	What work should I put into the house? Will it be worth it?
Real Estate Agents	Support your client: seller or buyer. Maximize your commission.	
Buyers	Negotiate with confidence!	

Let's Talk About Ames, IA

Ames is a city in the state of Iowa.

- Population: **66,427**
- Average Commute: **16 min**
- Median Household Income: **$48,105**

36 parks, trails, an aquatic center, and ice skating rink

IG: @cityofames

More on Ames: Major Employers

- Iowa State University (10K+)
- Iowa Department of Transportation (2-5k)
- Mary Greeley Medical Center (1300)
- City of Ames (1-2k)
- 3M (250-500)

IG: @iowastateu

@iowadot

@marygreeley100

The Dataset

Overview | Data | Code | Discussion | Leaderboard | Rules | Team | My Submissions | **Submit Predictions** | ...

Data Description

File descriptions

- **train.csv** - the training set
- **test.csv** - the test set
- **data_description.txt** - full description of each column, originally prepared by Dean De Cock but lightly edited to match the column names used here
- **sample_submission.csv** - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

Data fields

Here's a brief version of what you'll find in the data description file.

- **SalePrice** - the property's sale price in dollars. This is the target variable that you're trying to predict.

www.kaggle.com/

ML Procedure

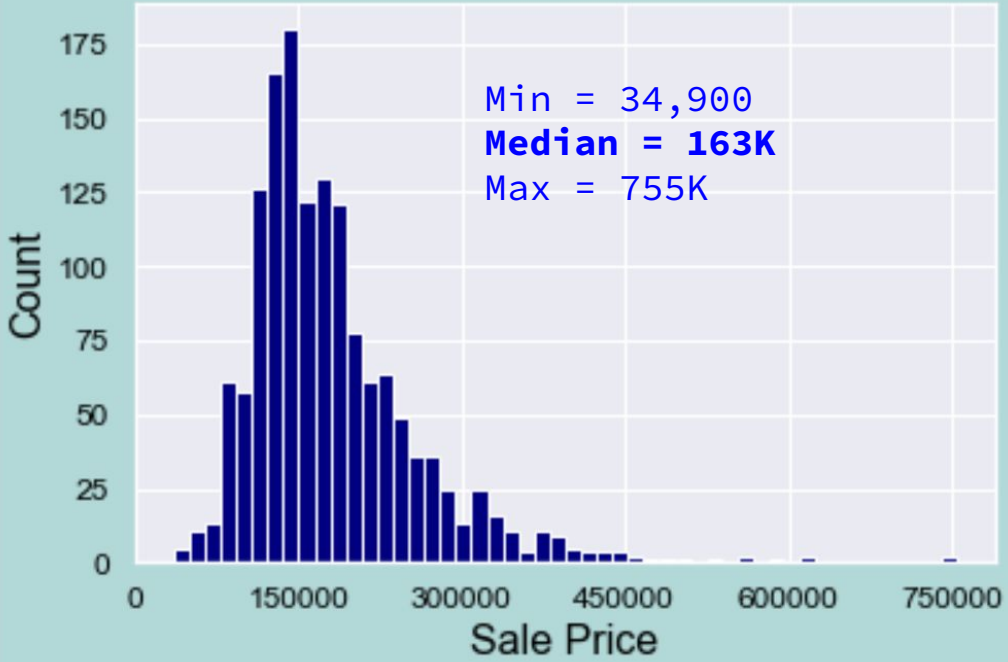
1. EXPLORATORY DATA ANALYSIS

- Load Dataset
- Examine Data

1460 rows

81 columns

Dependent Variable

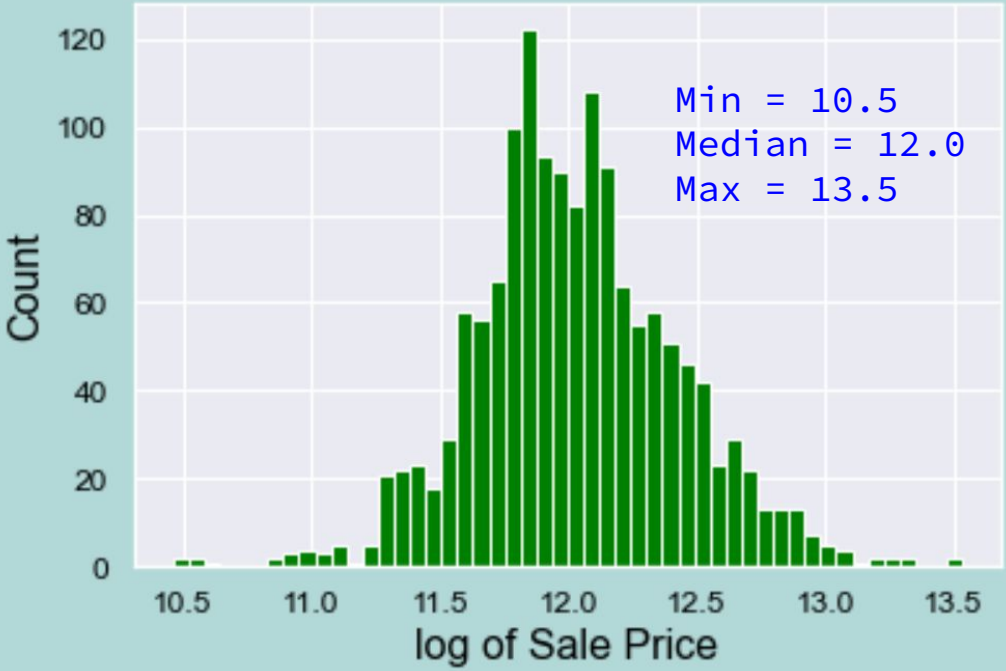


ML Procedure

1. EXPLORATORY DATA ANALYSIS

- Load Dataset
- Examine Data

Dependent Variable

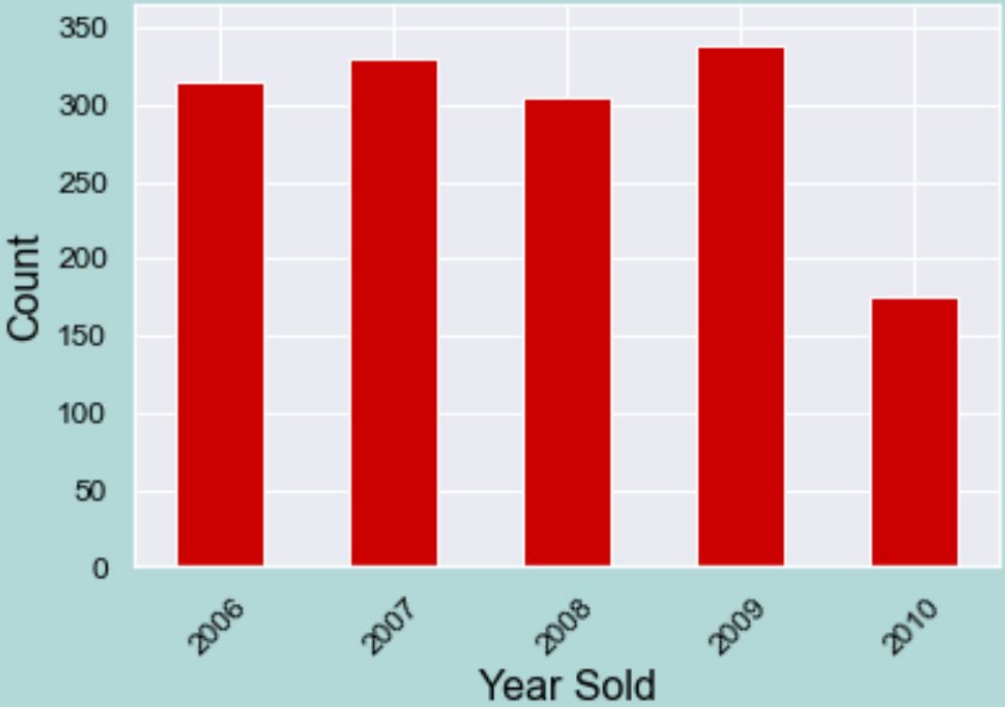


ML Procedure

1. EXPLORATORY DATA ANALYSIS

- Examine Data

Independent Variables



ML PROCEDURE

1. EXPLORATORY DATA ANALYSIS

15 Homes Removed

- Examine Data, looking for discrepancies, outliers

Removed 6 homes because the Total Bedrooms Above Ground = 0

Removed 1 home because Full Baths =0 (above ground or basement)

Removed 1 home because Kitchen Above Ground = 0, but there was a Kitchen Quality entry = Typical

Removed 1 home because Basement Finish Type 2 is missing, but Basement Finish Type 2 Square Footage != 0

Removed 5 homes because Masonry Veneer Type = None, but Masonry Veneer Area !=0

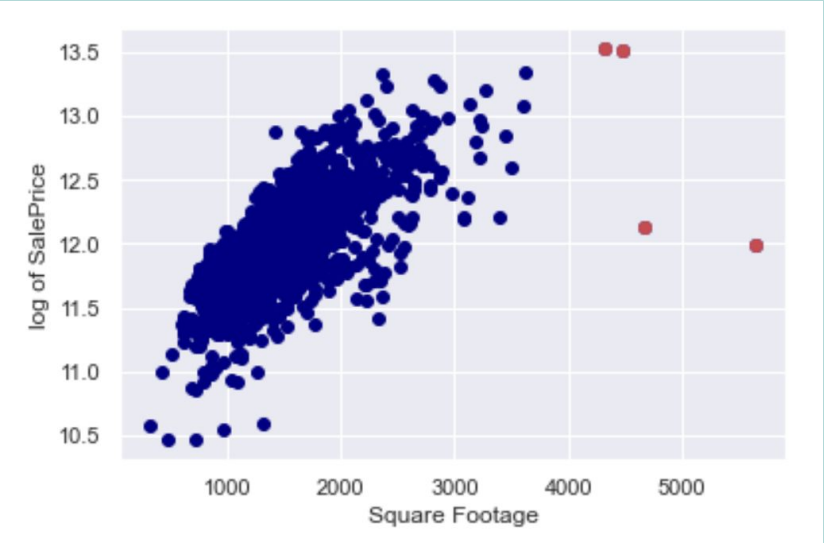
Removed 1 home because it had no electrical information

ML Procedure

1. EXPLORATORY DATA ANALYSIS

4 additional Homes Removed

- Examine Data, looking for discrepancies, outliers



Removed 4 homes because Square Footage (above ground) are outliers

Z-scores > 4.5

ML Procedure

1. EXPLORATORY DATA ANALYSIS

- Impute missing values

  -LotFrontage
  -MasVnrArea

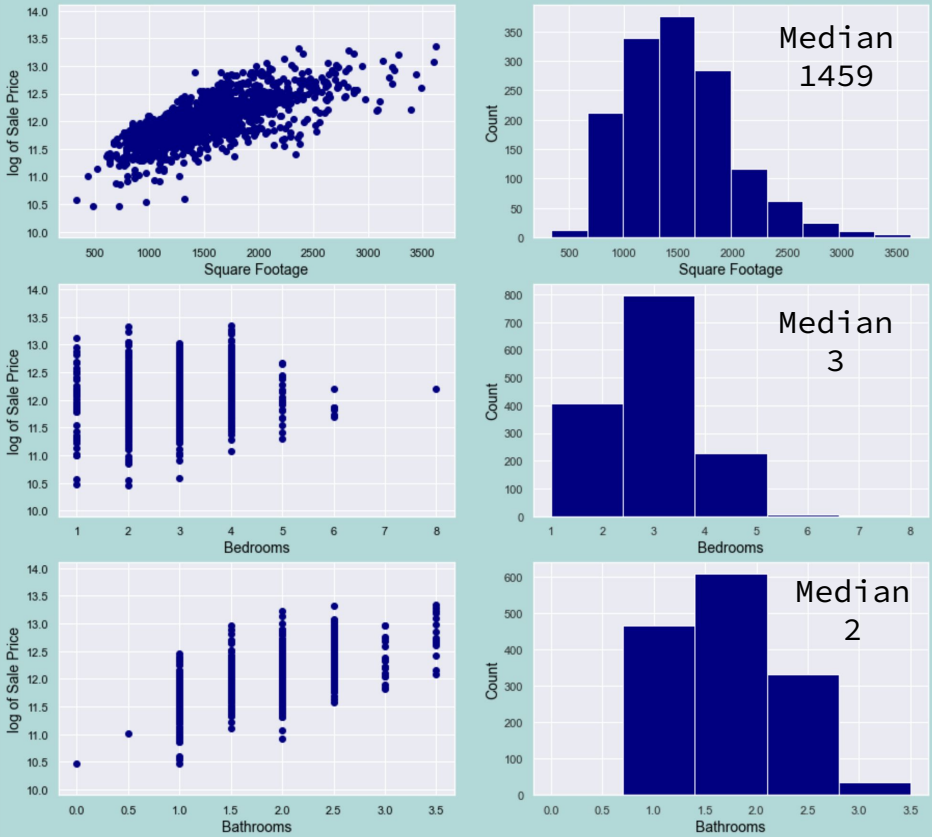
  -Bsmt variables
  -Garage variables
  -FireplaceQu
  -Pool variables
  -Alley
  -Fence

ML Procedure

1. EXPLORATORY DATA ANALYSIS

- Examine Data

Home Size

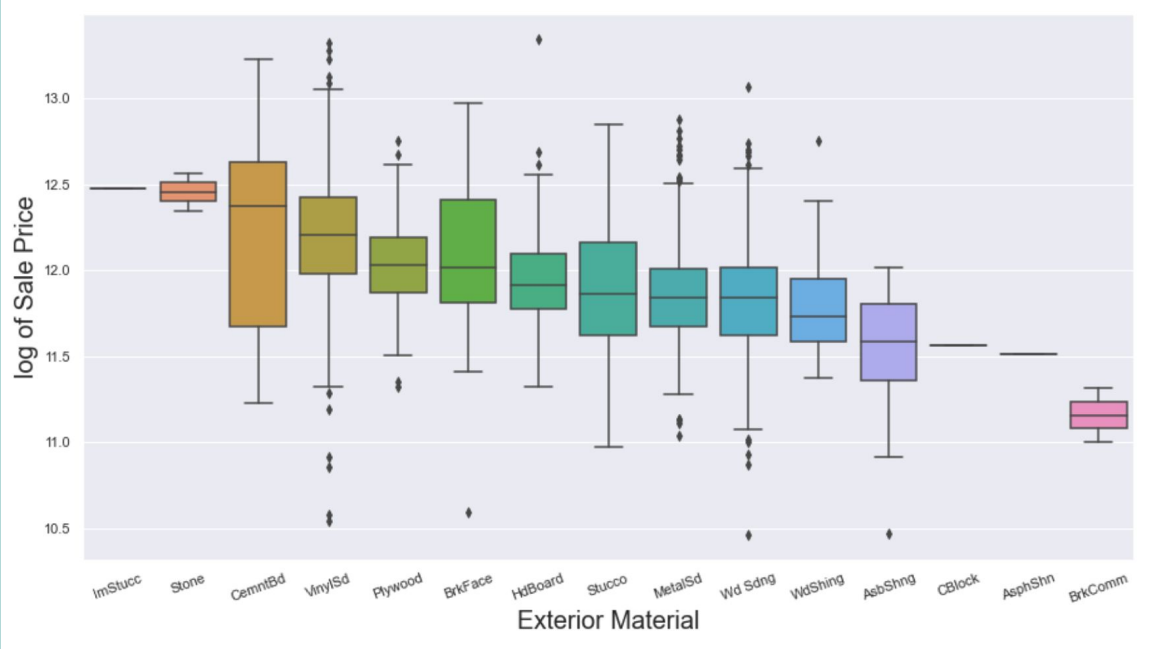


ML Procedure

1. EXPLORATORY DATA ANALYSIS

- Examine Data

Curb Appeal

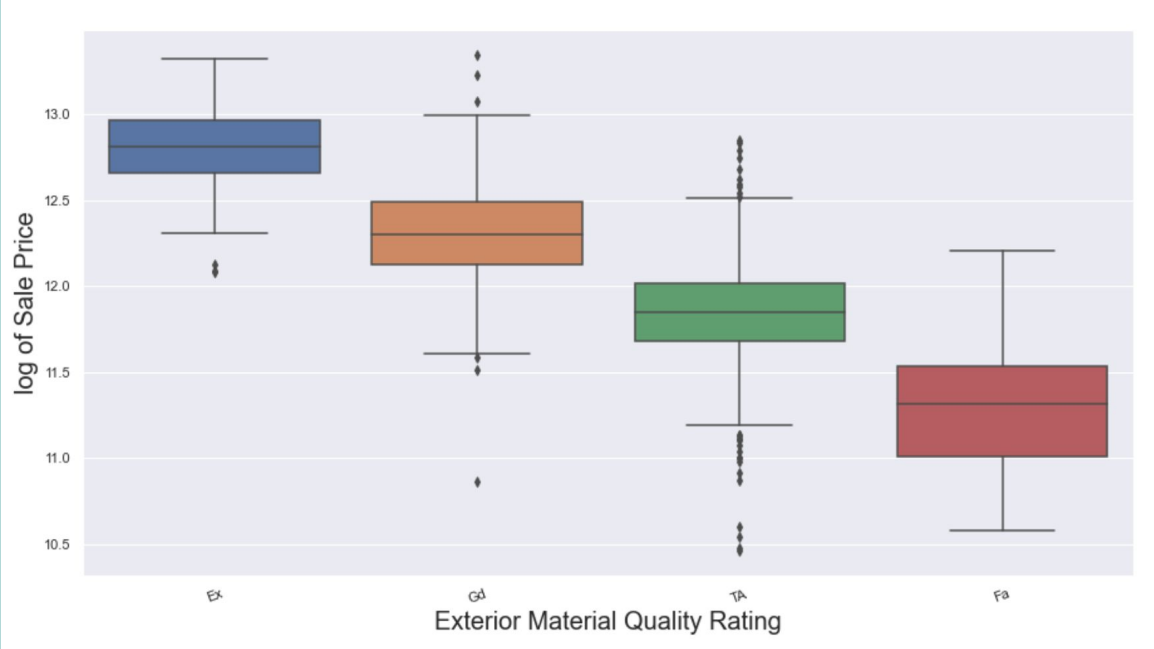


ML Procedure

1. EXPLORATORY DATA ANALYSIS

- Examine Data

Curb Appeal

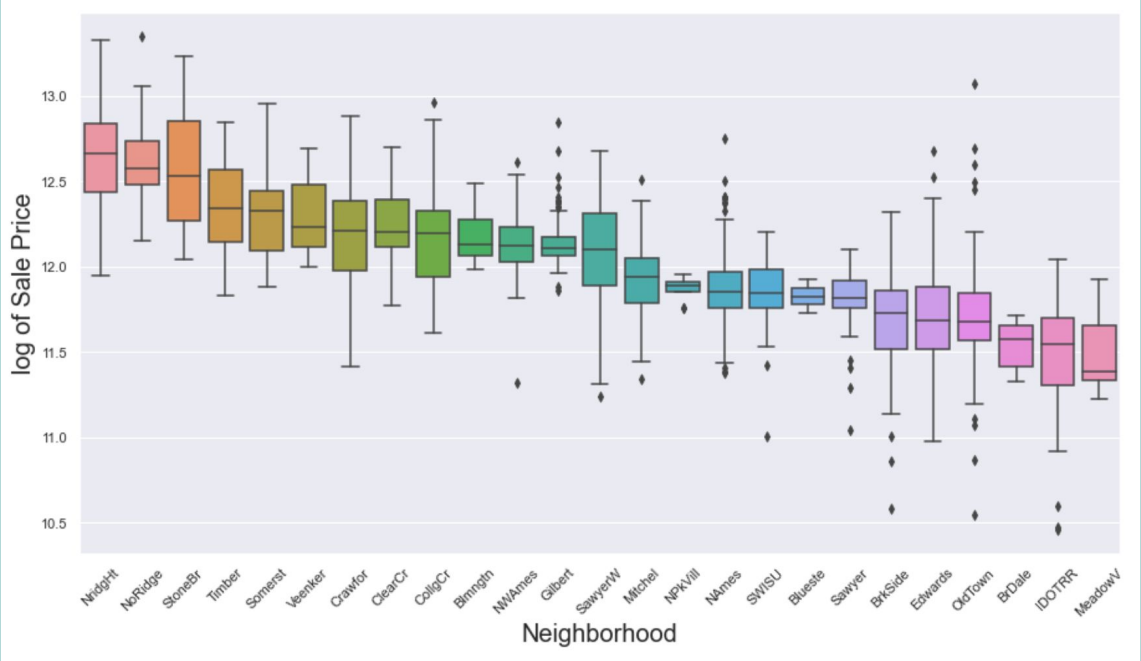


ML Procedure

1. EXPLORATORY DATA ANALYSIS

- Examine Data

Location

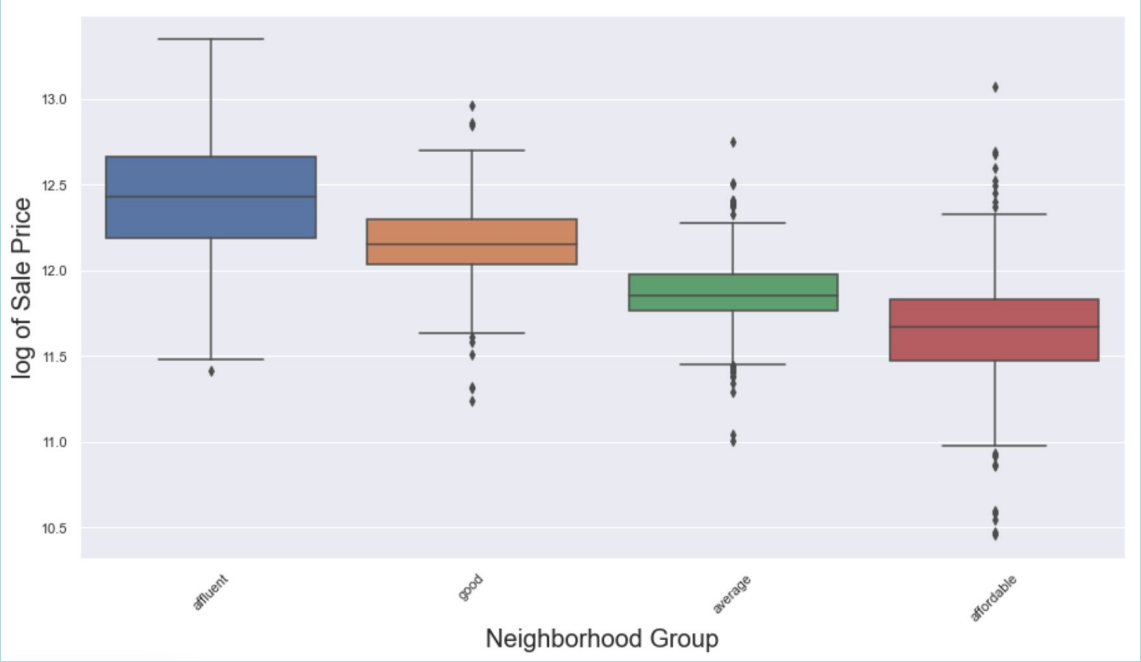


ML Procedure

1. EXPLORATORY DATA ANALYSIS

- Examine Data

Location



ML Procedure

2. Feature Engineering

- Create new Features

  AgeHome = Year Sold - Year Built
  YrsSinceRemodel = Year Sold - Year Remodeled
  Num_Bathrms_AbvGrd = FullBath + HalfBath*0.5

  Neighborhood_Grp
  hasGarage
  hasBasement

ML Procedure

2. Feature Engineering

- Select Features
  - My goal was to reduce the number of features that I sent into the Model training step.

Numerical Variables

Metric - correlation between log of Sale Price and the independent variable.

Correlation threshold = 0.15

22 (out of 30) Numerical variables met the criteria

ML Procedure

2. Feature Engineering

- Select Features
  - My goal was to reduce the number of features that I sent into the Model training step.

Categorical Variables

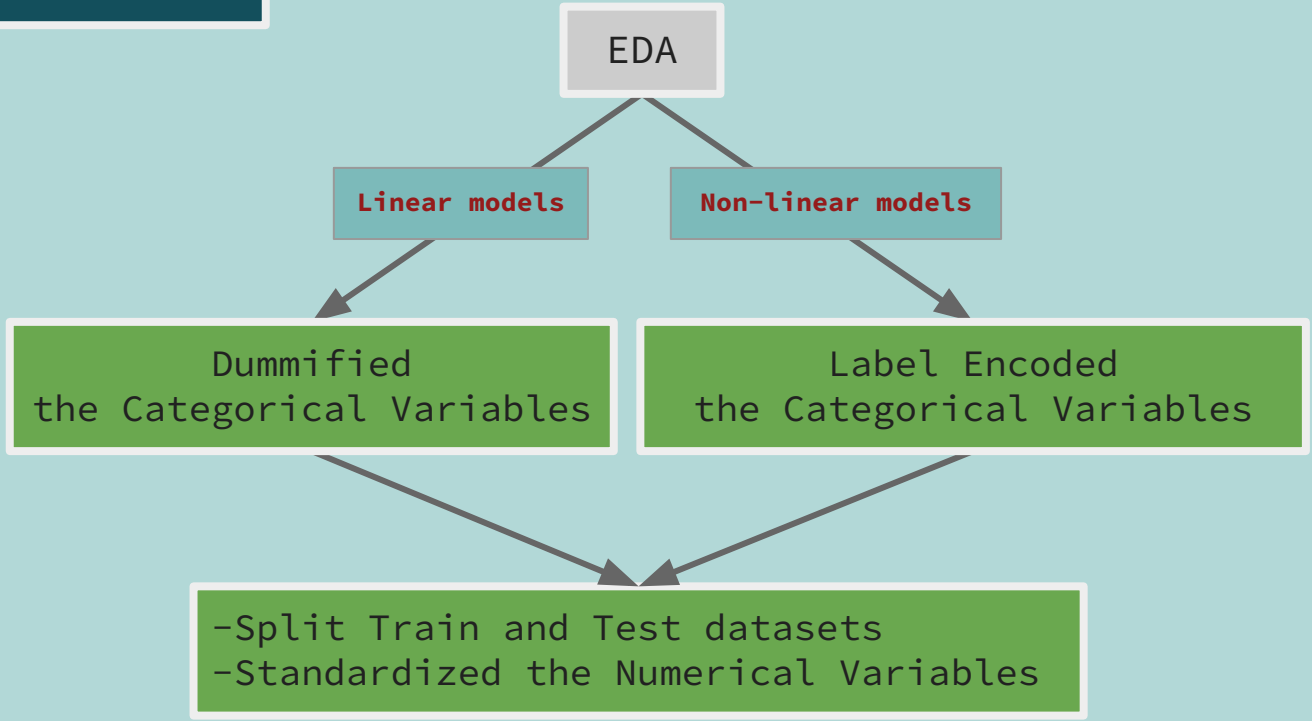
Procedure - Anova and effects size.

Anova p-value < 0.01
Effect size > 0.059

27 (49) categorical variables met the criteria

ML Procedure

3. Preprocessing



ML PROCEDURE

4. Train the Models

Model	Type	Hyperparameters
Linear Regression	Linear	None
Lasso	Linear	λ = 0.001
Ridge	Linear	λ = 0.01
Elastic Net	Linear	λ = 0.006 l1_ratio = 0.1
Random Forests	Non-linear	max_features = 6 n_estimators = 500 min_samples_split = 2
Gradient Boosting	Non-linear	learning_rate = 0.01 n_estimators = 2000 max_features = 4 min_samples_split = 2

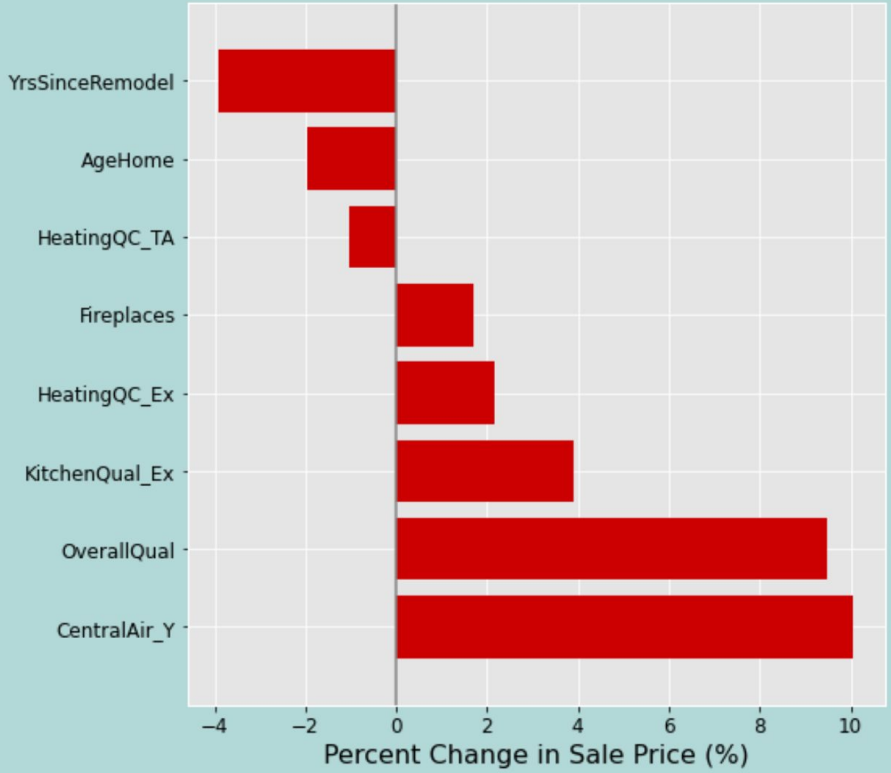
RESULTS

Evaluate Models

Model	RMSE Logarithmic Train	RMSE Logarithmic Test	RMSE (dollars) Train	RMSE (dollars) Test	r^2 Train	r^2 Test
Elastic Net	**0.1162**	**0.1156**	**20,786**	**19,013**	**0.92**	**0.91**
Lasso	0.1175	0.1159	21,073	18,943	0.91	0.90
Ridge	0.1057	0.1198	19,509	18,846	0.93	0.90
Random Forests	0.0526	0.1272	10,378	20,689	0.98	0.89
Gradient Boosting	0.0830	0.1281	14,110	21,137	0.96	0.88
Linear Regression	0.1063	nan	19,621	nan	0.93	nan

Take-Aways

House Features

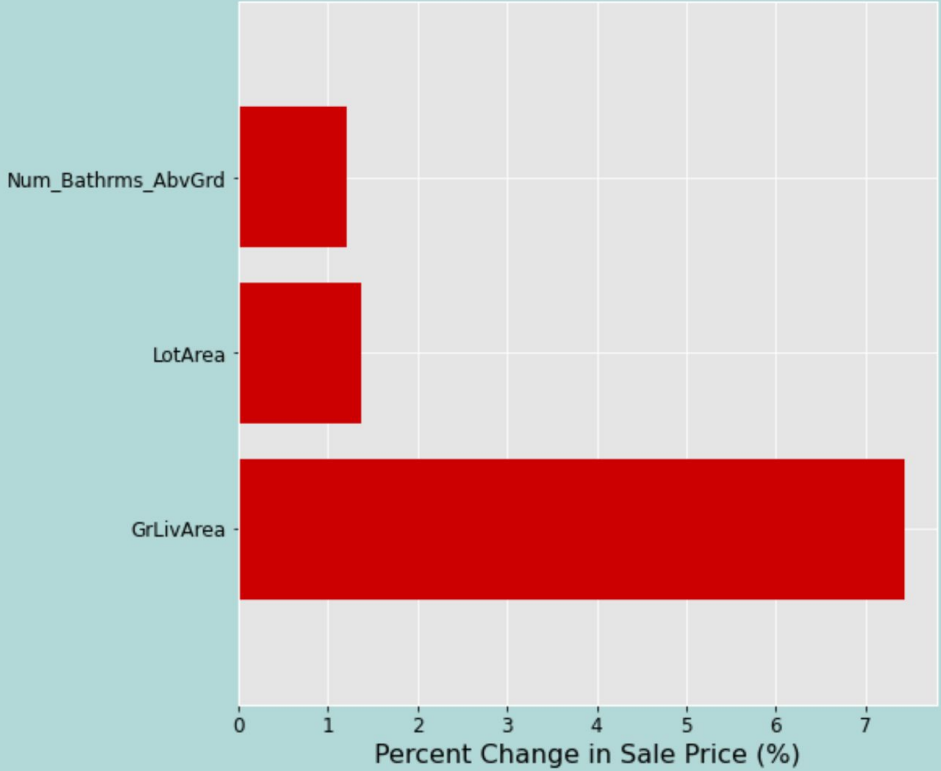


Increase in Sale Price:

- Central Air
- **High Quality Kitchen**
- **High Quality Heating System**
- Fireplace

Take-Aways

Home Size



Increase in Sale Price:

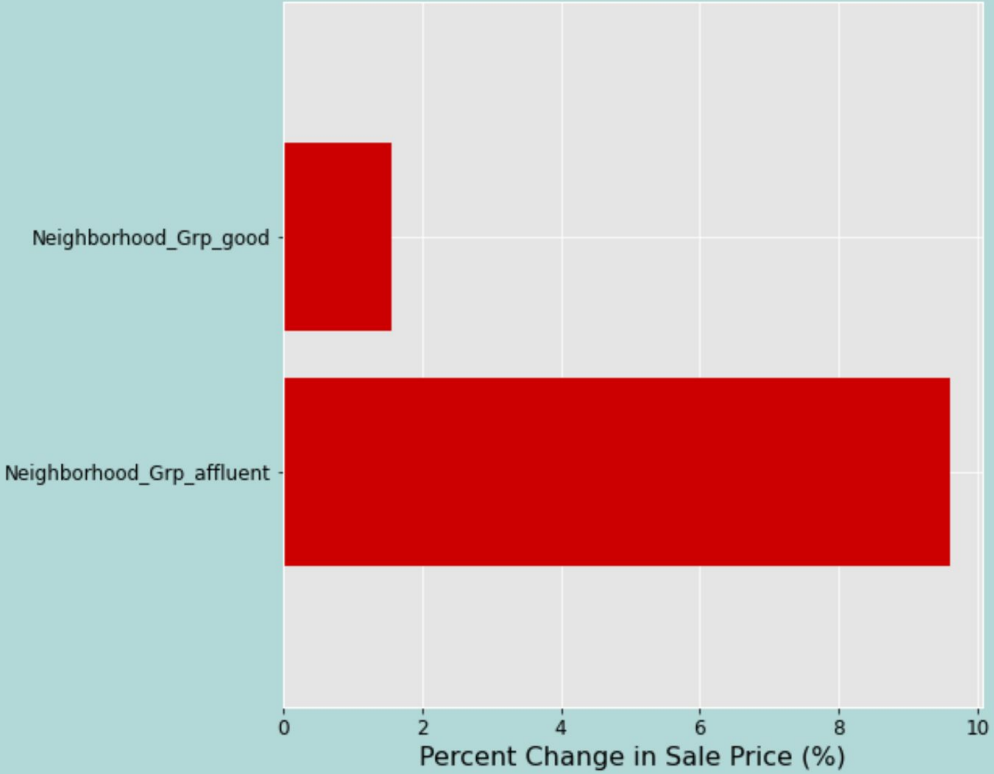
- Home Square Footage
- Number of Bathrooms

An increase in 500 square feet could result in an Sale Price increase of 7%.

An increase in 1 bathroom could result in a Sale Price increase of 1%.

Take-Aways

Location



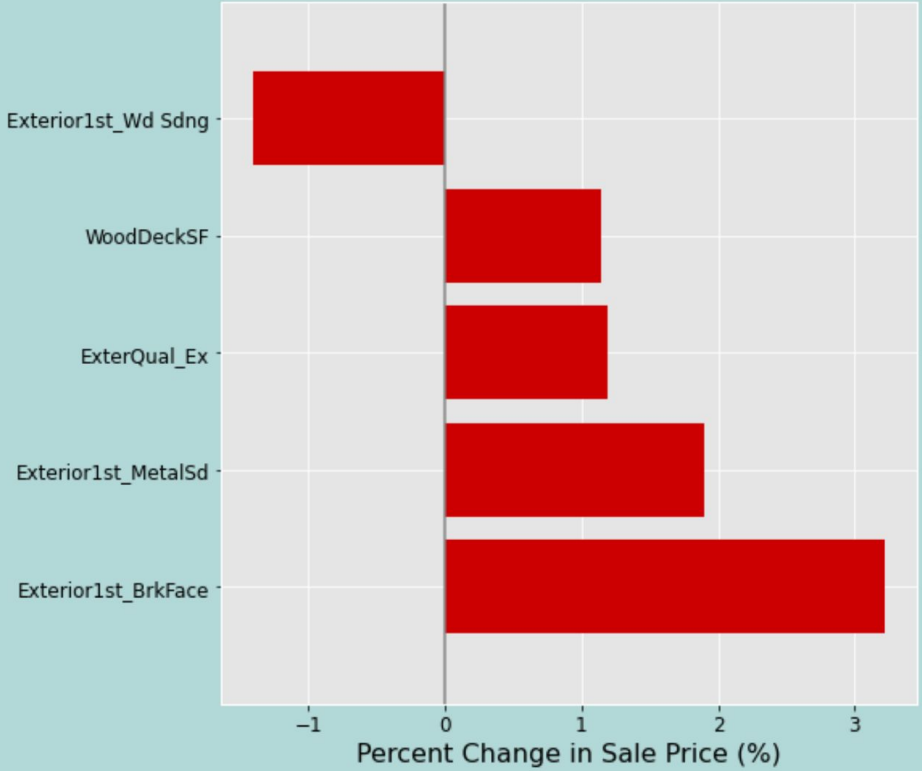
<u>Increase</u> in Sale Price:

- Homes in the following neighborhoods:

  NridgHt
  NoRidge
  StoneBr
  Timber
  Somerst
  Veenker
  Crawfor

Take-Aways

Exterior

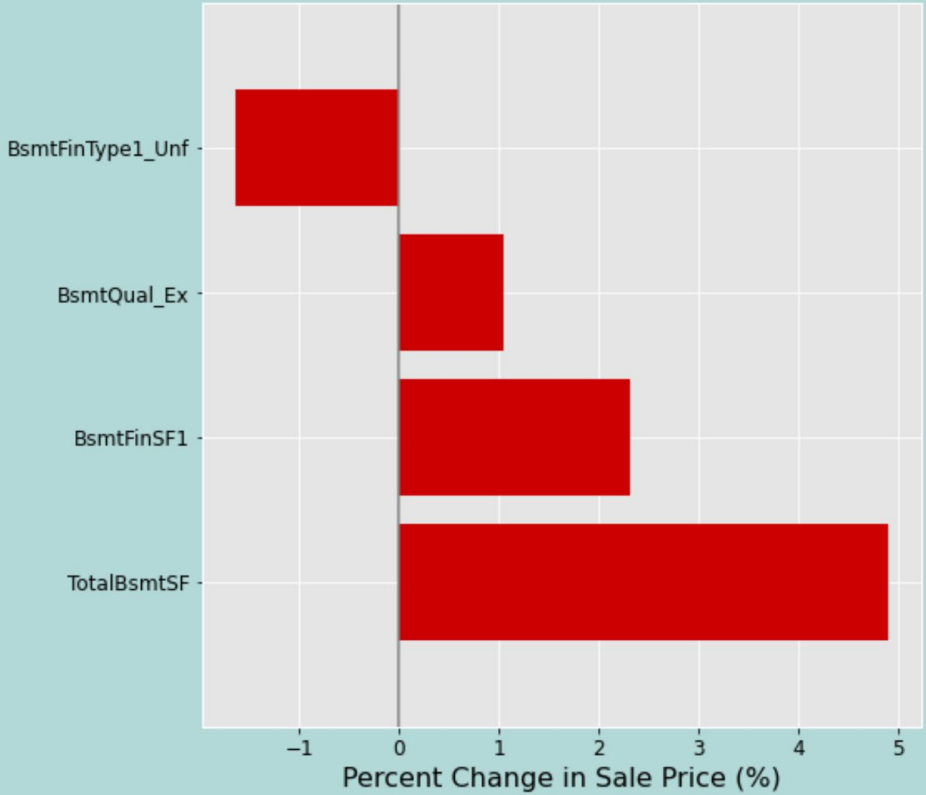


Increase in Sale Price:

- Exterior Brick Face
- Exterior Metal Siding
- **Excellent Exterior Quality**
- **Wood Deck Square Footage**

Take-Aways

Basement

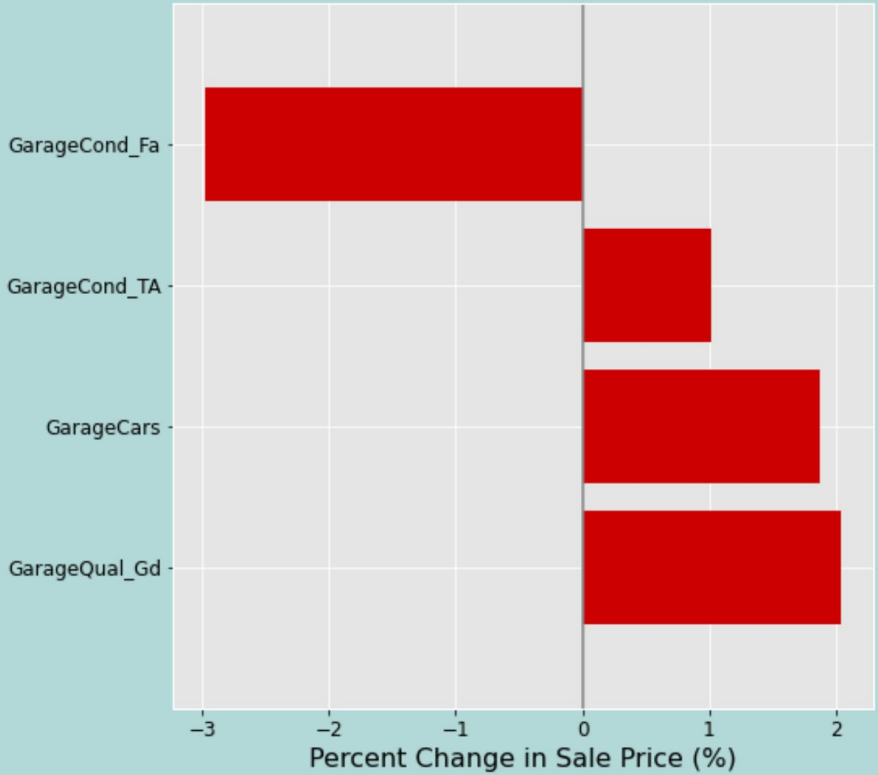


Increase in Sale Price:

- Larger Basement
- **Larger Finished Basement**
- Excellent Quality Basement

Take-Aways

Garage



Increase in Sale Price:

- Good or Average Garage Quality
- Larger Garage

Conclusions

Seller: Things to consider

- **Update Heating System (2%)**
- **Update your Kitchen (4%)**
- **Add a Bathroom (1%)**
- **Improve the Exterior Quality (1%)**
- **Finish the Basement**

My Future Work: Look into where the model is failing and flag homes for further analysis.

Goal: expand EDA and improve both the Feature Engineering and Feature Selection steps.

Thank You!

~~For~~

Sale