Modeling Ames, Iowa Real Estate

GA DSI - Project 2

Emily Siegel

PROBLEM STATEMENT

There are certain aspects about a home that you simply can't change like lot size, age of the home, and square footage for the most part. Of course, there are endless features that you can add, fix or upgrade. As a homeowner, it can be overwhelming trying to evaluate the ROI for home improvement projects.

This project not only aims to build a model for predicting sale prices for homes in Ames, Iowa, but also tries to determine if there are particular home qualities that have the potential to help increase a home's sale price for when you're ready to sell it.

01

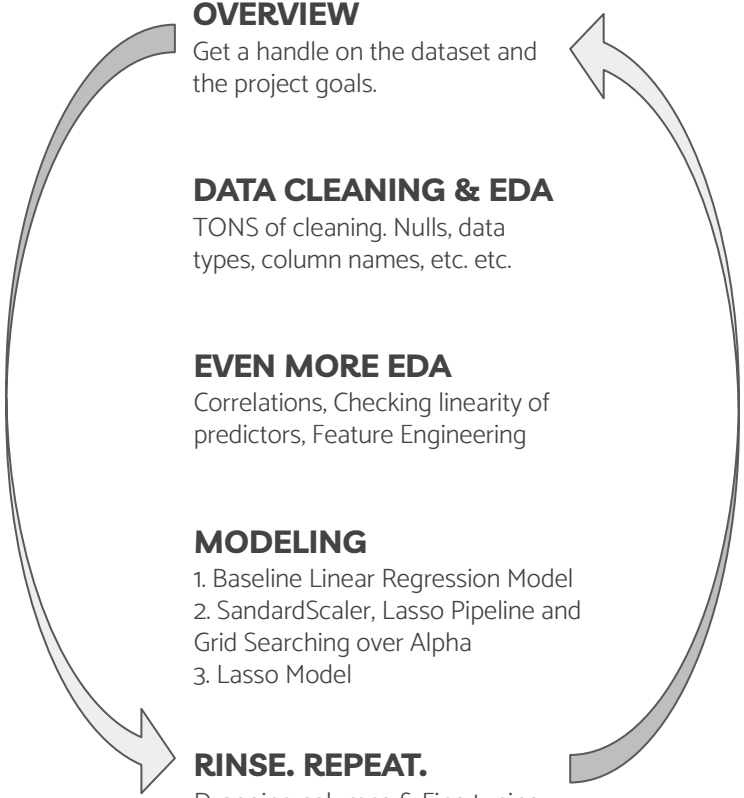
02

03

04

05

MODELING WORKFLOW



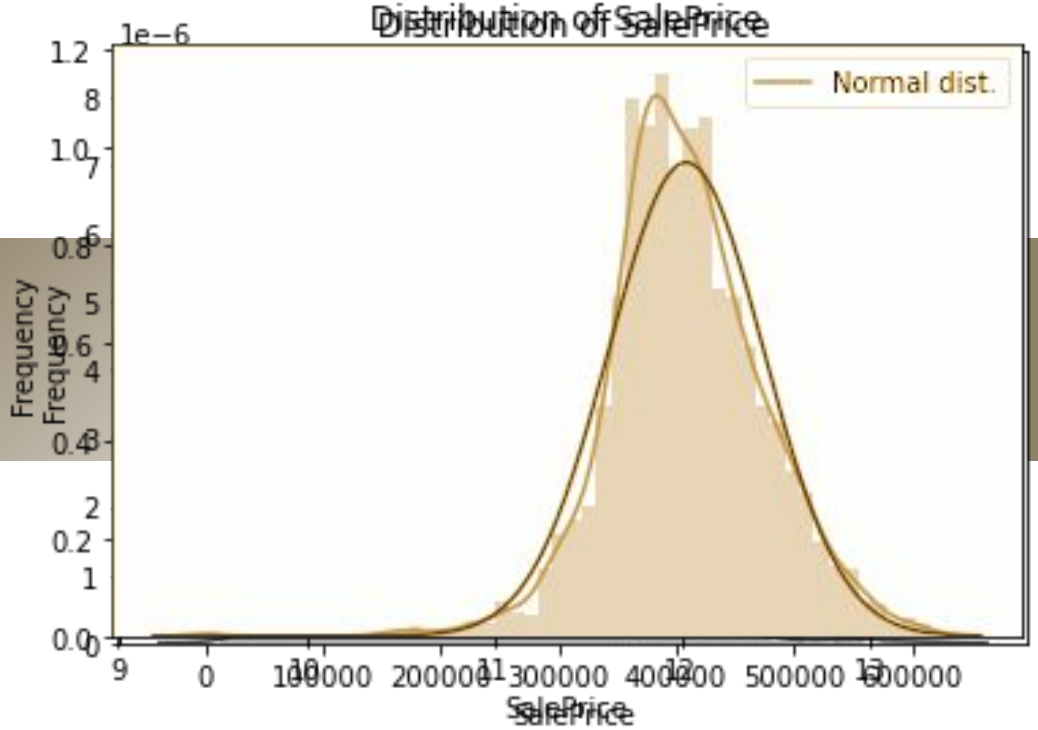
OVERVIEW
Get a handle on the dataset and the project goals.

DATA CLEANING & EDA
TONS of cleaning. Nulls, data types, column names, etc. etc.

EVEN MORE EDA
Correlations, Checking linearity of predictors, Feature Engineering

MODELING
1. Baseline Linear Regression Model
2. SandardScaler, Lasso Pipeline and Grid Searching over Alpha
3. Lasso Model

RINSE. REPEAT.
Dropping columns & Fine tuning model parameters and scaling features if needed.



TARGET: SALE PRICE

Applied a log transformation

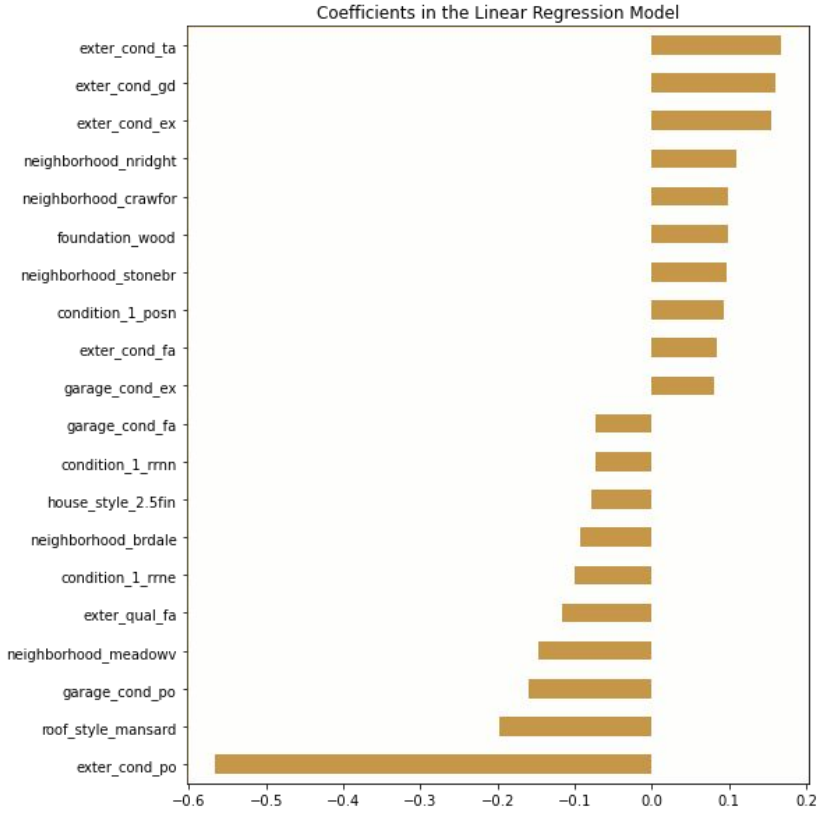
FEATURE ENGINEERING

NUMERIC FEATURES

- Total SqFt
- House Age
- Total Bathrooms

CATEGORICAL FEATURES

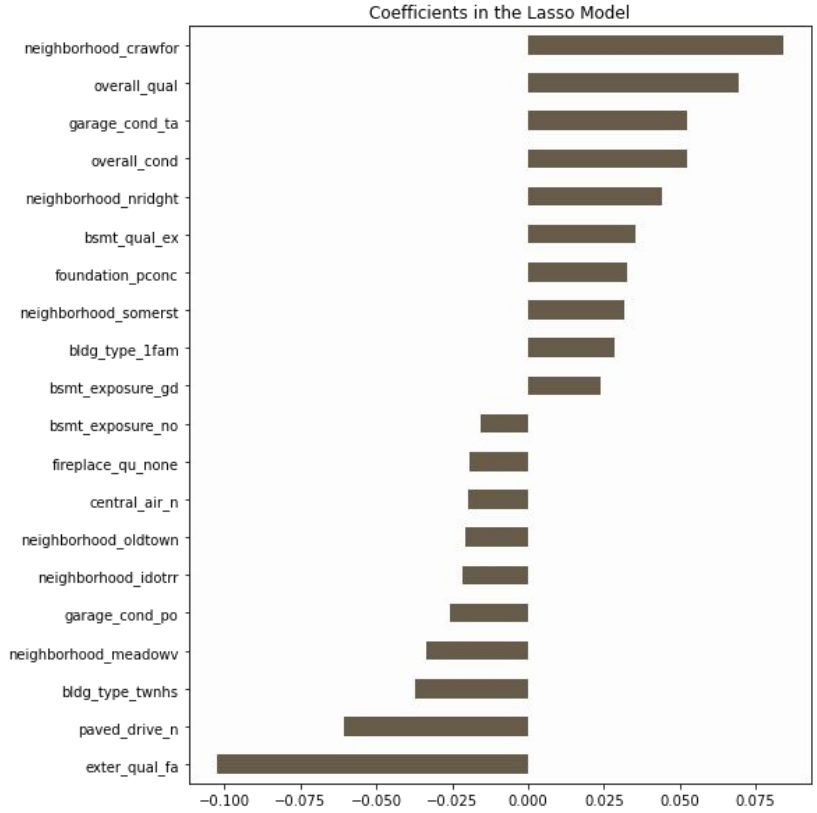
- Dummified
- Dropped certain categories not contained in the test dataset



Overfit model - too many predictors and maybe need to engineer different ones!

BASELINE LINEAR REGRESSION MODEL

TRAIN R2	93.22%
TEST R2	89.66%
RMSE	0.1306



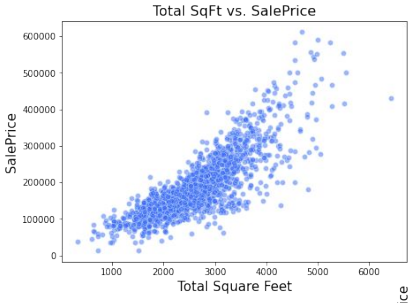
Lasso picked 69 variables and eliminated the other 106 variables

PRODUCTION LASSO MODEL

TRAIN R2	91.76%
TEST R2	90.54%
RMSE	0.1249

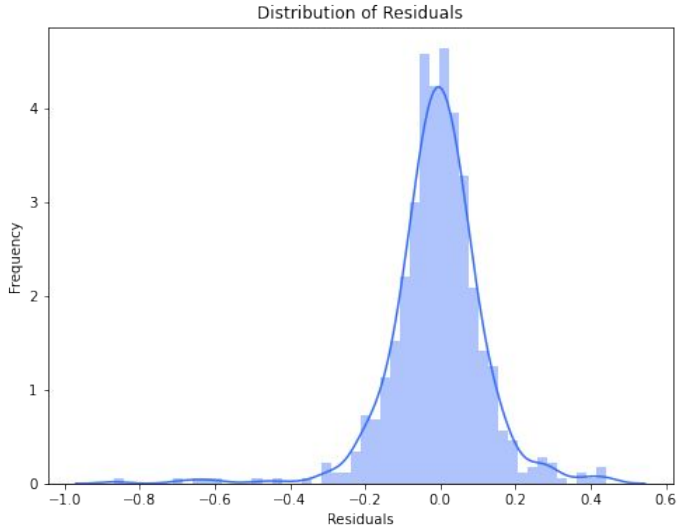
ACTUAL vs. PREDICTIONS

MODEL EVALUATION



LINEARITY?

Linear relationship between the predictors and the response variable

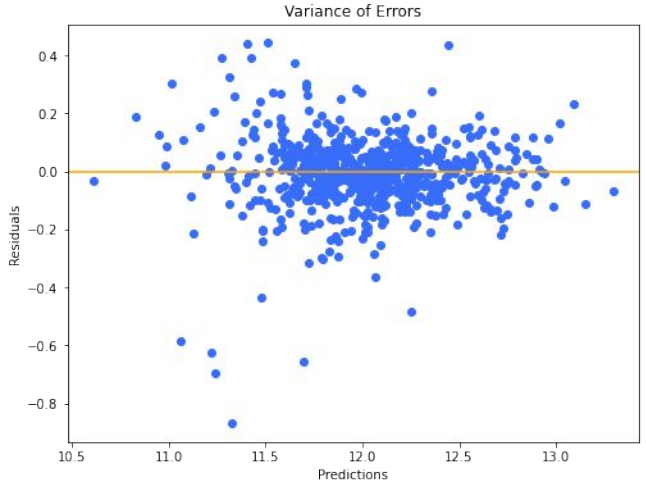


NORMALITY OF ERRORS?

Some skewness but relatively normally distributed

INDEPENDENT ERRORS?

This one is assumed...



MODEL EVALUATION

EQUAL VARIANCE OF ERRORS?

These look pretty equal, but a few outliers...

NO MULTICOLLINEARITY?

Independence of Predictor Variables - Dropped some probable overlap in predictors. I suspect there are still some collinear features I need to deal with.

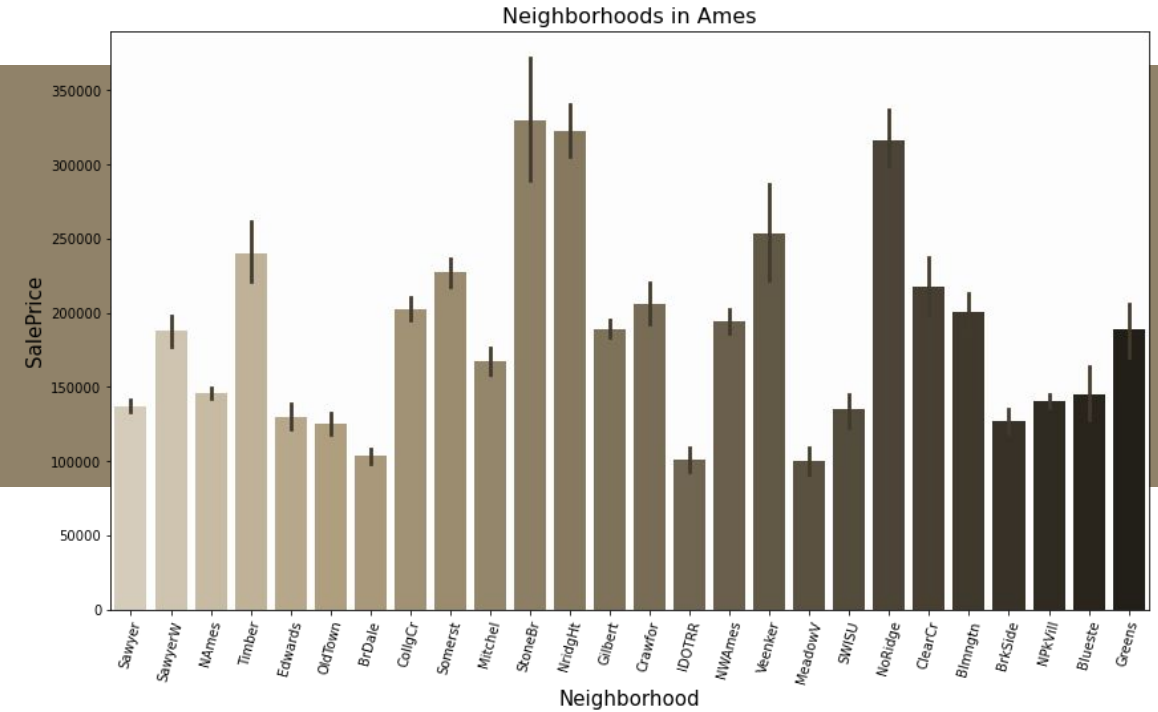
- GarageType: Garage location
  - 2Types More than one type of garage
  - Attchd Attached to home
  - Basment Basement Garage
  - BuiltIn Built-In (Garage part of house - typically has room above gar
  - CarPort Car Port
  - Detchd Detached from home
  - NA No Garage
- GarageYrBlt: Year garage was built
- GarageFinish: Inte[illegible] finish of the garage
  - Fin Finished
  - RFn Rough Finished
  - Unf Unfinished
  - NA No Garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
  - Ex Excellent
  - Gd Good
  - TA Typical/Average
  - Fa Fair
  - Po Poor
  - NA No Garage
- GarageCond: Garage condition
  - Ex Excellent
  - Gd Good

ALL GARAGE!

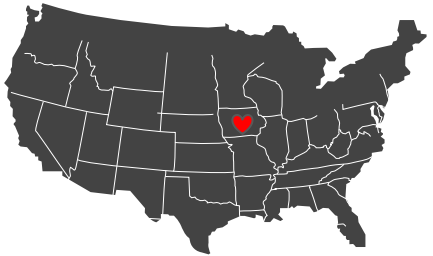
- BsmtExposure: Walkout or garden level basement walls
  - Gd Good Exposure
  - Av Average Exposure (split levels or foyers typically score average or above)
  - Mn Mimimum Exposure
  - No No Exposure
  - NA No Basement
- BsmtFinType1: Quality of basement finished area
  - GLQ Good Living Quarters
  - ALQ Average Living Quarters
  - BLQ Below Average Living Quarters
  - Rec Average Rec Room
  - LwQ Low Quality
  - Unf Unfinished
  - NA No Basement
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
  - GLQ Good Living Quarters
  - ALQ Average Living Quarters
  - BLQ Below Average Living Quarters
  - Rec Average Rec Room
  - LwQ Low Quality
  - Unf Unfinished
  - NA No Basement
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area

ALL BSMT!

SALE PRICE vs. NEIGHBORHOOD



1. Stone Brook
2. Northridge Heights
3. Northridge



POTENTIALLY HIGHER ROI HOUSE PROJECTS?

POOL AREA

For every 1 sqft increase in Pool Area, Sale Price can be expected to increase by holding all else equal

WOOD DECK SF

For every 1 sqft increase in wood deck area, Sale Price can be expected to increase holding all else equal

BSMT FIN SF 1

For every 1 sqft increase of a Finished Basement, Sale Price can be expected to increase holding all else equal

CENTRAL AIR

For every 1 unit of Central Air added, Sale Price can be expected to increase holding all else equal

Overall Quality has a large impact!

QUESTIONS?

REFERENCES

1. http://jse.amstat.org/v19n3/decock/DataDocumentation.txt
2. Pardoe , I. (2008), “Modeling home prices using realtor data”, Journal of Statistics Education Volume 16, Number 2 (2008). http://jse.amstat.org/v16n2/datasets.pardoe.pdf
3. https://www.kaggle.com/fedi1996/house-prices-data-cleaning-viz-and-modeling
4. Presentation template by Slidesgo