AMES HOUSING

Modeling Sale Price

Kade Higgins

PROBLEM STATEMENT

To identify the best model for predicting Sales Prices in Ames, Iowa.

ABOUT THE DATA

Housing Data in Ames, IA

Descriptive Info related to Property

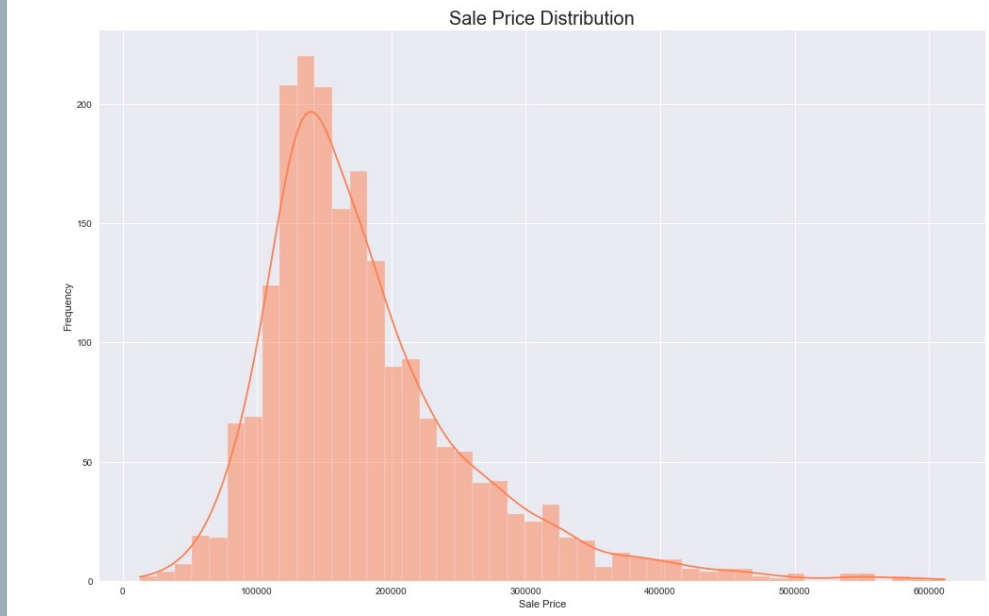
Ex. Overall Qual, Neighborhood, Fireplaces, Lot Shape, Yr Sold, etc

Null Values

Replaced with N/A, 0 as appropriate

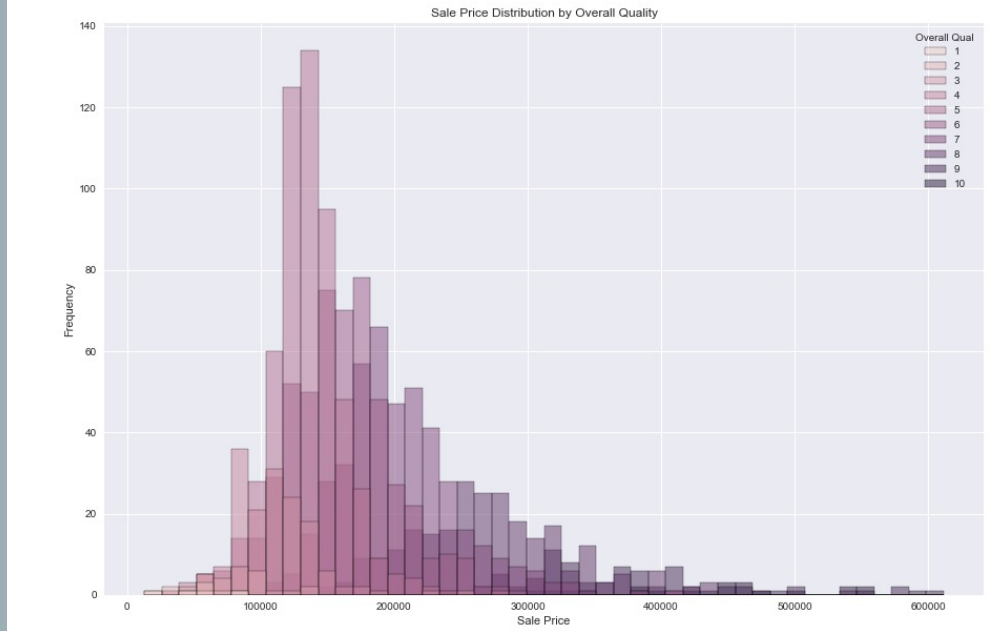
EXPLORATORY DATA ANALYSIS

SALE PRICE DISTRIBUTION



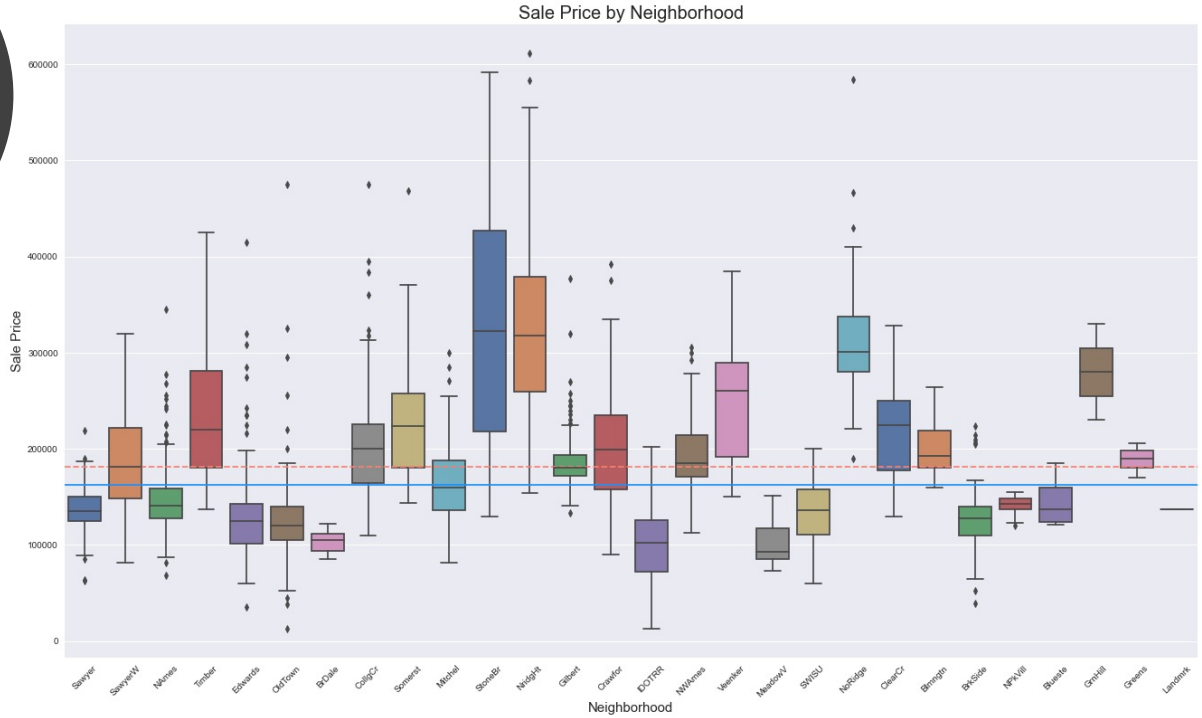
EXPLORATORY DATA ANALYSIS

SALE PRICE
X
OVERALL QUAL



EDA
SALE PRICE
X
LIVING AREA

EDA
SALE PRICE
N.HOOD
PLOTS



MODEL 1: LINEAR REGRESSION

A simple univariate Linear Regression using 'Overall Qual' as the X variable:

- 66.2% Accuracy
- RMSE: 45571.38
- Y = 44318.64X - 89719.7
- *our baseline

MODEL 2 LOGISTIC REGRESSION

- Three Variables (Overall Qual * Exter Qual, Gr Liv Area, Neighborhood)
- OHE on Neighborhood
- Standard Scaler + Logistic Regression
- 0.58% Accuracy
- RMSE = 44139.14

MODEL 2.5 RIDGE

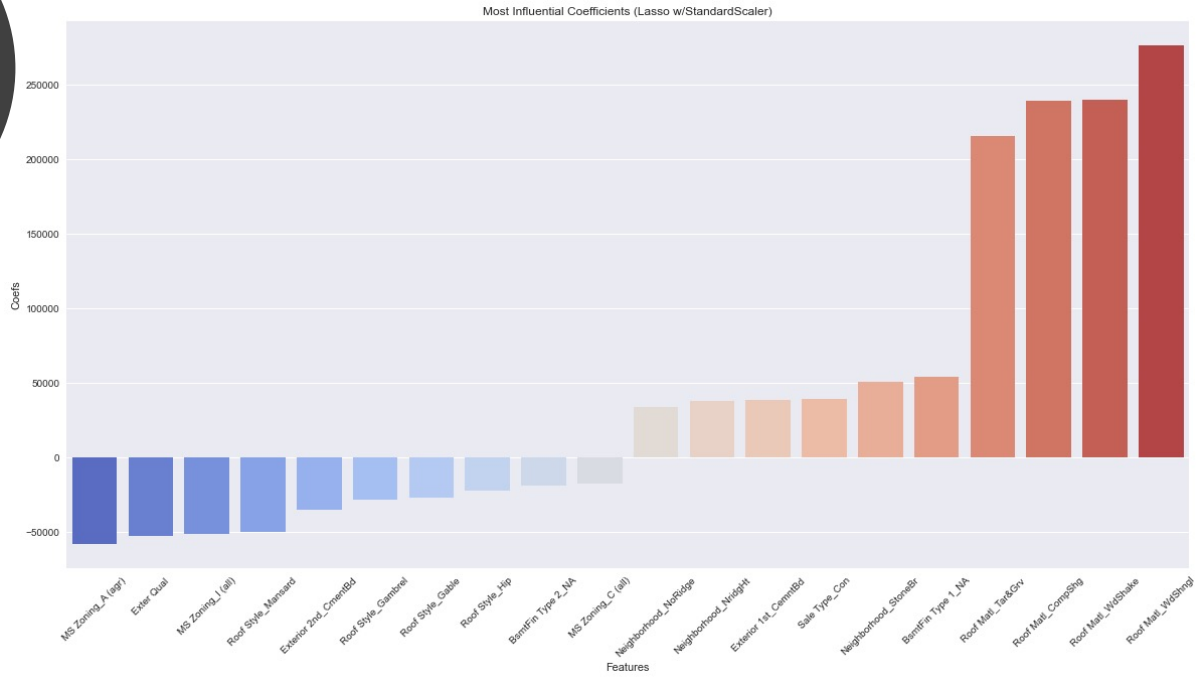
- Three Variables (Overall Qual * Exter Qual, Gr Liv Area, Neighborhood)
- Swapped *Logistic Regression* out for ***Ridge***
- Best Ridge Alpha: **20**
- 81.4% Accuracy
- RMSE = 33989.4

MODEL 3:
A LITTLE OF EVERYTHING

This model keeps nearly all features, uses a pipeline to do OHE, StandardScaler, SelectKBest and Lasso

- ***90.3% Accuracy***
- ***RMSE*** *24685.86*
- Best Alpha (10), chose 150 features

PLOTTING MODEL 3 COEFS



CONCLUSIONS

Complex Model Outperforms Simpler Ones

But, not by much.

- Location is likely a ***very*** strong predictor of SalePrice, particularly in combination with variables related to quality

Next Time…

- Look at different models with a few variables each
- Our model 3 wasn't a huge gain over model 2.5, despite using nearly all variables compared to 3
- Investigate collinearity, multicollinearity, possibly log transform the target

THANK YOU FOR LISTENING

I'm happy to answer any questions

Kade Higgins