Predicting the Sale Price of an Ames, IA Home

Contents

Problem Statement

Can we construct a model that accurately predicts sale price given numeric and categorical data for a home in Ames, IA?

EDA and Data Cleaning

1. Dataset was obtained from Kaggle
2. https://www.kaggle.com/c/dsi-us-11-project-2-regression-challenge/

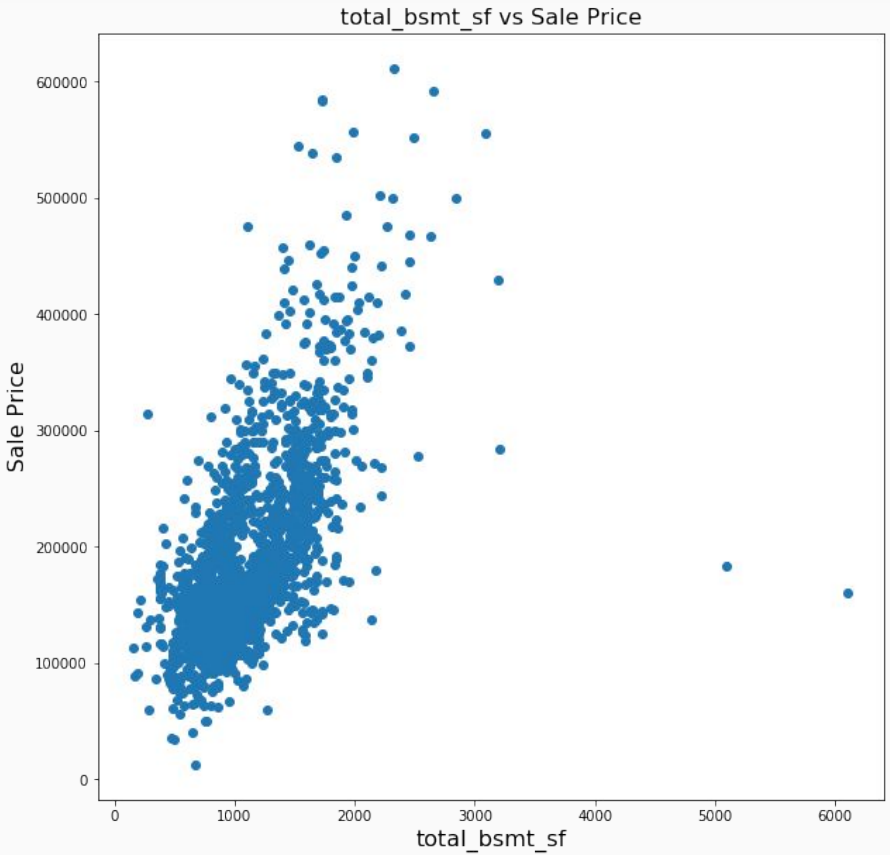
EDA and Data Cleaning

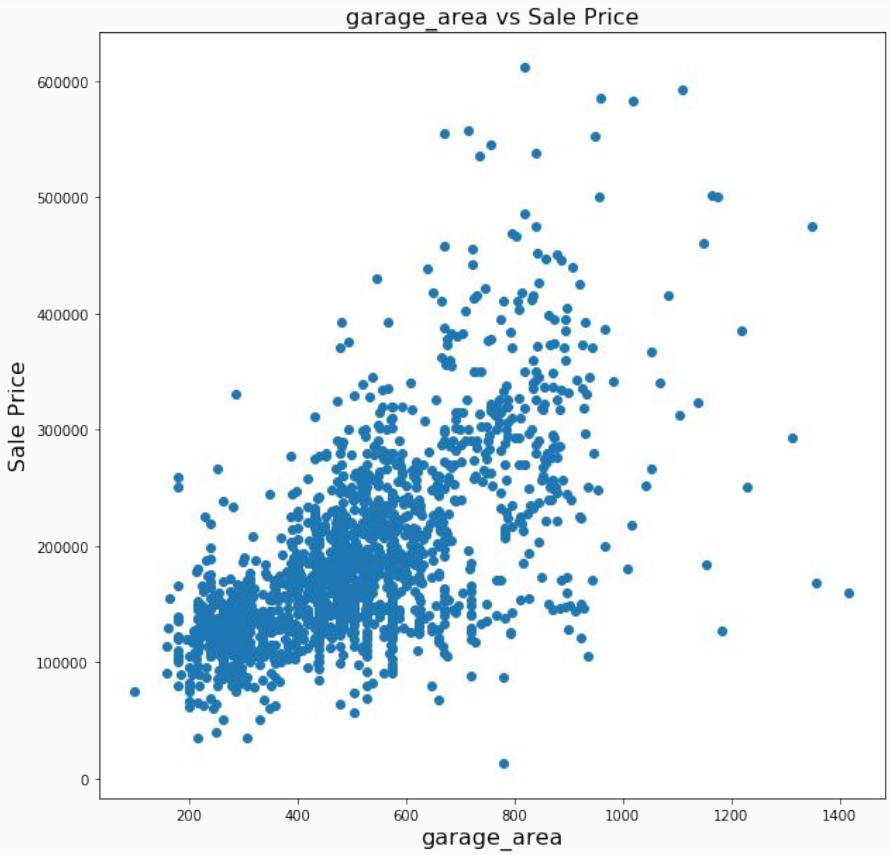
1. Original dataset has about 80 Features
2. Details about these features can be found here:

EDA and Data Cleaning

Some features were useful...

EDA and Data Cleaning

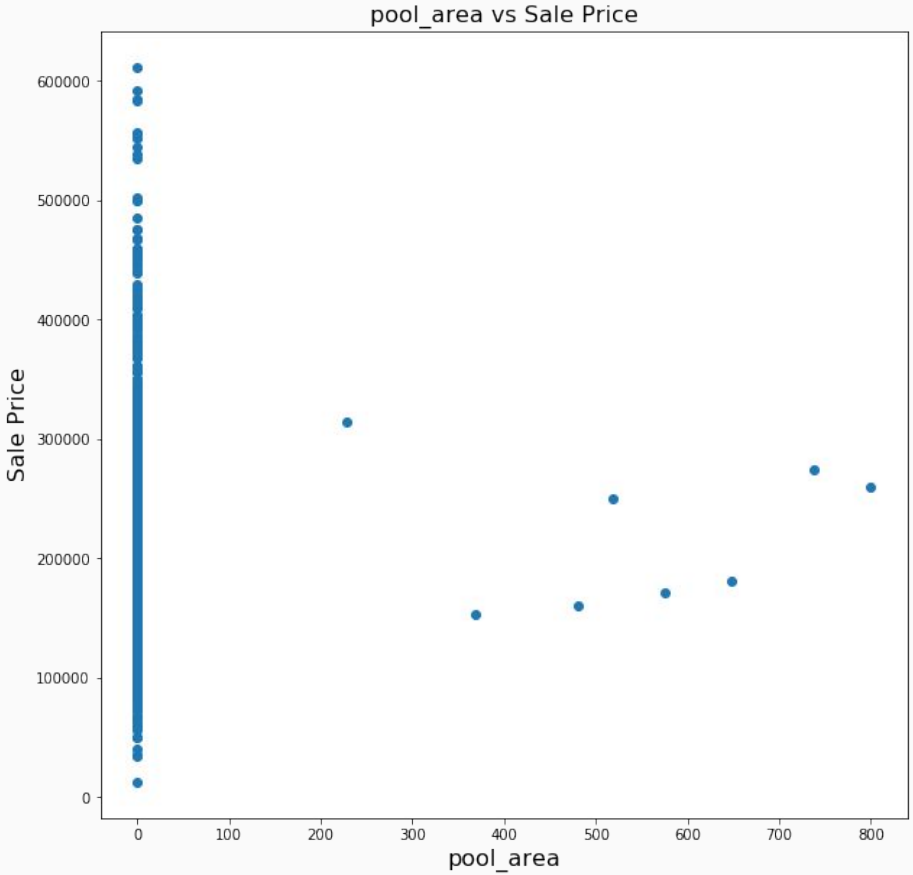


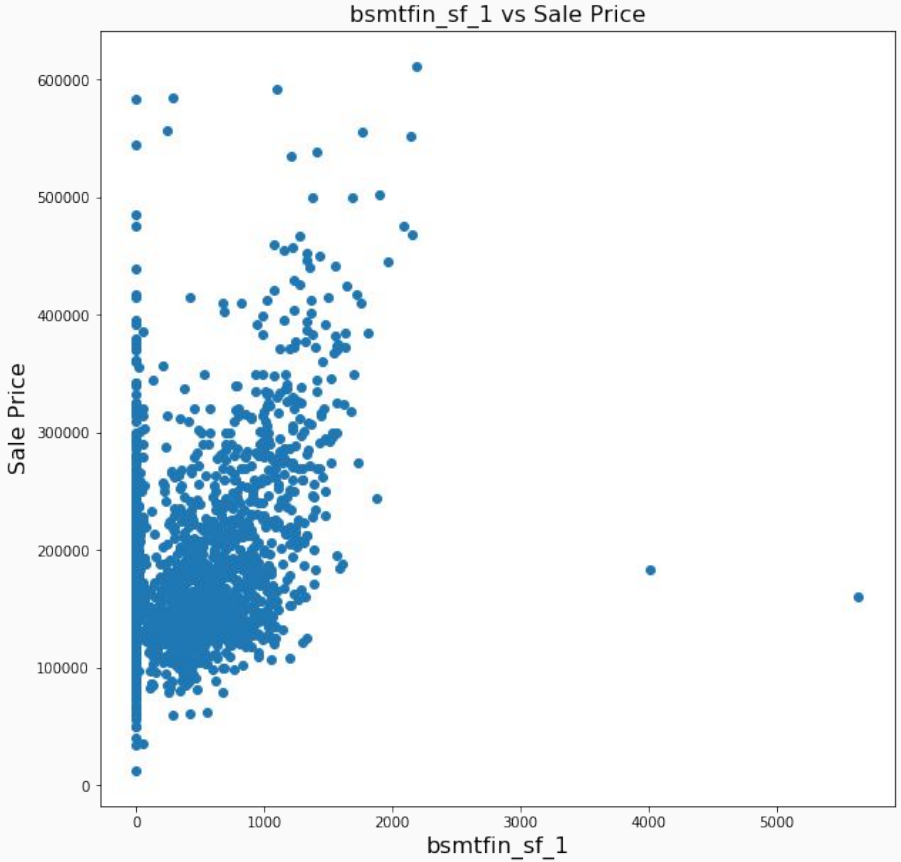


EDA and Data Cleaning

Others were maybe less so...

EDA and Data Cleaning





EDA and Data Cleaning

1. Numeric Data
    a. Only high (> 50%) correlation features were used
2. Categorical Data
    a. Features that were heavily weighted in one value were removed
    b. E.g. ‘basement condition’ was 92% a single value, it was removed
3. Missing values were imputed using Sklearn’s SimpleImputer
    a. ‘Most frequent’ strategy was used

Modeling Procedure - Preprocessing

1. Train-test-split
2. Preprocessing Pipeline
    a. Standard Scaler
    b. One Hot Encoding

Modeling - Regression Models

1. Linear Regression
2. k-Nearest Neigbors
3. Ridge (L2-norm)
4. LASSO (L1-norm)

Modeling - Evaluation Metrics

1. R2
2. Mean Absolute Error
3. Mean Squared Error

Modeling

K-Nearest Neighbors Proved to be the most effective model:

1. Train R2 ~ 87%
2. Test R2 ~ 83%
3. Next best was LASSO with Train and Test R2 respectively 86% and 82%

Modeling

1. KNN model
2. Ideal world, all data points lie on straight line
3. That would be perfect prediction of data

Hyperparameter Tuning

1. GridSearchCV was used to tune the knn model
2. Search Parameters:
    a. N_neighbors - range(1, 51, 10),
    b. Weights - ['uniform','distance'],
    c. P - [1,2]
3. Best Parameters
    a. N_neighbors - 11
    b. Weights - ‘distance
    c. P - 2

Hyperparameter Tuning

1. Train R2: 85%
2. Test R2: 84%

Recommendations

1. This model will likely generalize to other Midwestern U.S. cities provided:
    a. Comparable time period
    b. Comparable city size
2. To generalize beyond the midwest:
    a. Use aggregate climate and economic data instead of neighborhood

Thank your time!
Merci pour votre attention!