PREDICTING SALE PRICES OF HOMES IN AMES, IA

By: Cassy Clark

Objective

- Create a model that would predict the sale prices for homes in Ames, Iowa
- Given training data set, testing data set, and data dictionary
- Steps taken: EDA, Feature Engineering, Lasso, Ridge, Linear Regression
- Why should I care about Ames, IA?

Data Cleaning and EDA

- Train shape: (2051, 81) → (2051, 121)
- Test shape (878, 80) → (878, 119)
- Addressing null values
  - *Is it a missing value or does the house not have one?*
  - *What about lot frontage?*
  - *Any others?*
- New columns created: total bath, total sqft, total sqft*qual, fireplace*fireplace quality, beds*baths
- Dummy columns: central air, street, neighborhood, alley, garage type

Data cleaning cont.

```
#creating a function to switch our categorical data to numerical data for quality/condition
def qual_to_num(string):
    if string == 'Ex': #represents excellent
        return 5
    elif string == 'Gd': #represents good
        return 4
    elif string == 'TA': #represents typical/average
        return 3
    elif string == 'Fa': #represents fair
        return 2
    elif string == 'Po': #represents poor
        return 1
    else:
        return 0
```

- Global functions to transform categorical ranking to numerical ranking
- Used on both training and testing
- Difference between ordinal and Nominal
- Nominal data dummied

Feature Engineering

Created mini data frames to check for multicollinearity between different features and the target variable of sale price			
Outdoors	Garage	Neighborhood	Etc

Dropped columns in training set that were not in testing set	
Green Hills	Landmark

Variables created from feature engineering:			
Total sqft	Total sqft*quality	Fireplace*fireplace quality	Beds*baths

Models Used

- Multiple linear regression, lasso, ridge
- Train, test, split data
- Various feature combinations for independent variables
- Baseline RMSE: $79,239

First Round of Model Testing

Multiple Linear Regression

- Split 80/20
- R2 score of 0.830 and 0.862
- RMSE: $28,644

Lasso

- Same variables as linear regression model
- Multiple alphas used, best score I found was with the default alpha of 1
- R2 scores: 0.831 and 0.862
- RMSE: $28, 642

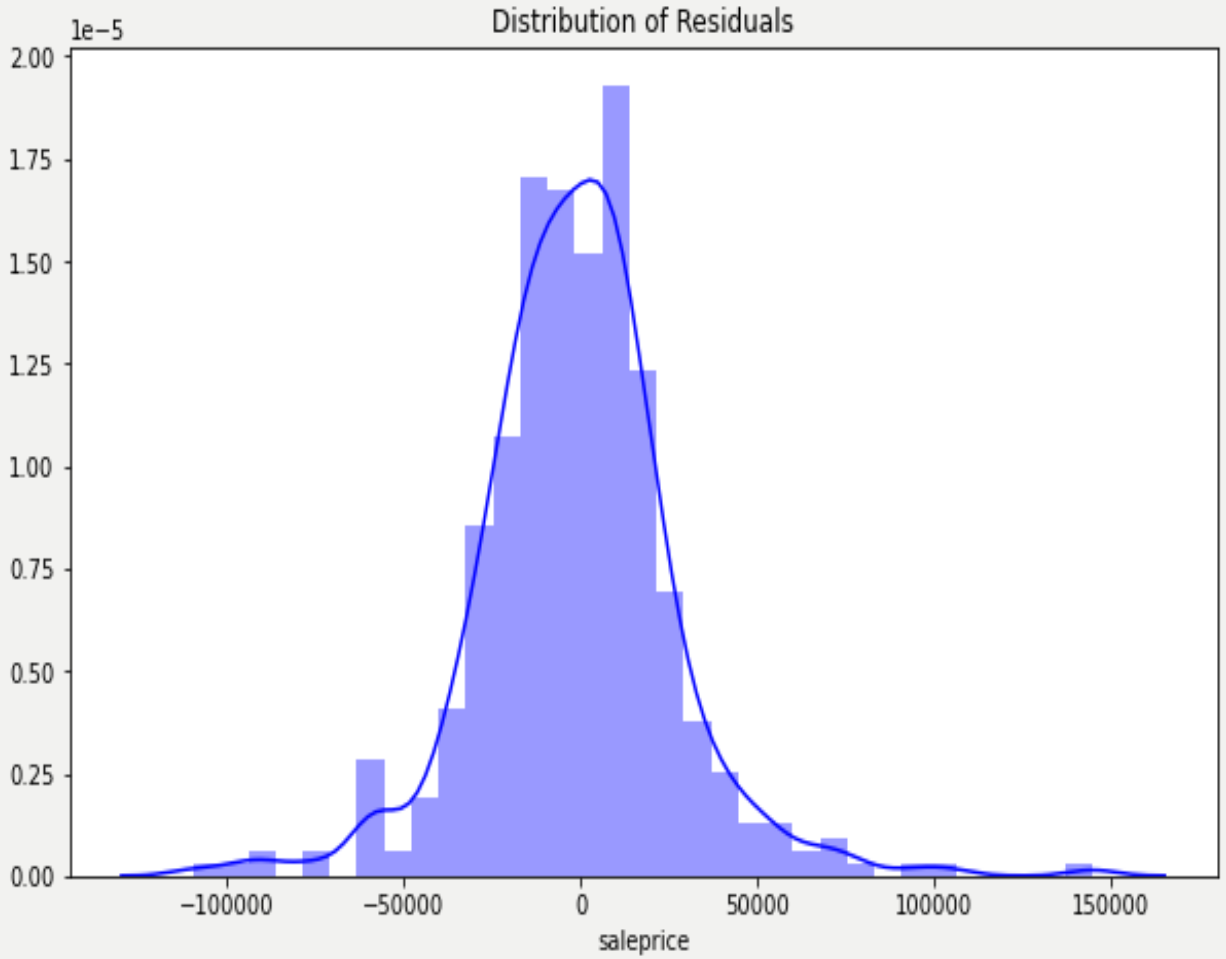
Ridge

- Same features as linear regression
- Best alpha: 100
- R2 scores: 0.829 and 0.865
- RMSE: $28,300

Best Model

- Multiple Linear Regression
- Split 80/20
- R2 score: 0.838 and 0.872
- RMSE: $27,599

Distribution of Residuals



Relationship between different features

- Overall quality: $7,414
- Total baths: $10,633
- North Ridge neighborhood: $39,366
- North Ridge Heights neighborhood: $42,857
- Stonebrook neighborhood: $60,703
- Exterior quality: $11,046
- Kitchen quality: $11,399

Conclusions

Best model recap

Model usage

Next steps