Modeling Home Prices in Ames, IA

Caitlin Streamer

Ames, IA is consistently rated as one of the best places to live in the US

Highlights:

- Population 66,000
- Iowa State University
- 30 miles north of Des Moines
- Home to small and big business
- Comprehensive medical care
- Parks and recreation
- Stable economy
- Emerging cultural scene

There is value in modeling home prices in Ames, IA, as the city continues to experience growth

Utilized dataset with 79 features on 2,051 homes sold in Ames from 2006 - 2010

Data science problem statement:

Develop a model to predict sale price of homes in Ames based on property features

Features at a glance:

- Sale Price
- Neighborhood
- Lot Area
- Building Type
- Year Built
- Overall Quality
- Bedrooms
- Fireplaces

- Garage Area
- Sale Type
- Fence
- Kitchens
- Full Bath

Performed substantial data cleaning and feature engineering to consolidate features

New Features Added

Name	Definition
Age	Year Sold - Year Remodeled
Is New	Year Sold == Year Built
Is Remodeled	Year Built == Year Remodeled
Total Bath	Sum of full bath and 0.5 x half bath per floor
Total SF	Sum of square footage per floor

Strongest Linear Relationships with Sale Price
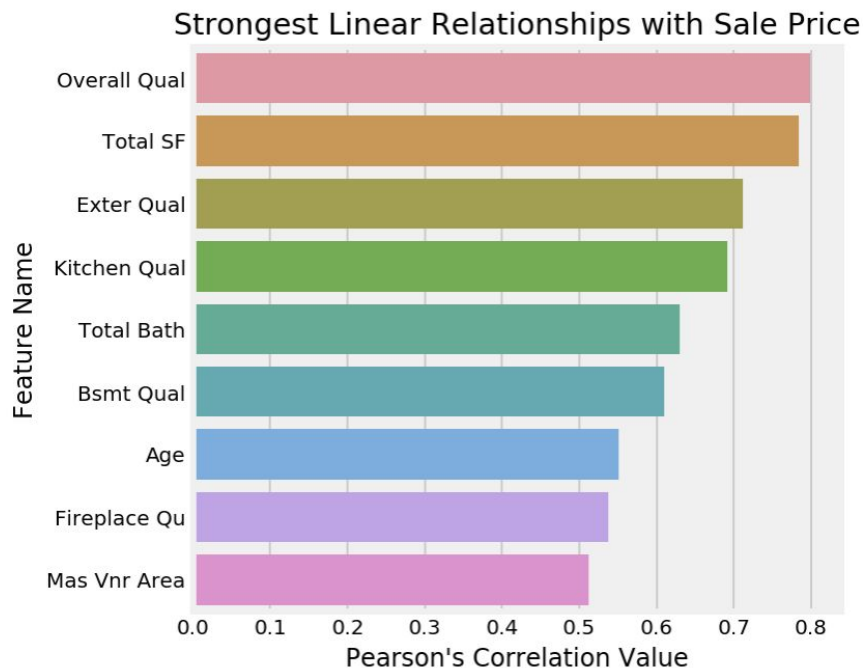
Overall Quality and Total Square Footage have the strongest linear correlation with Sale Price
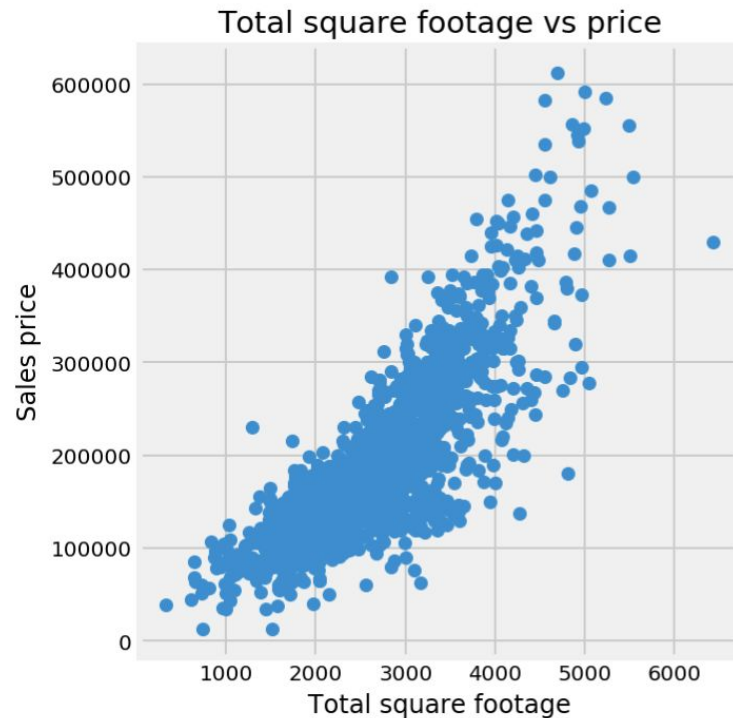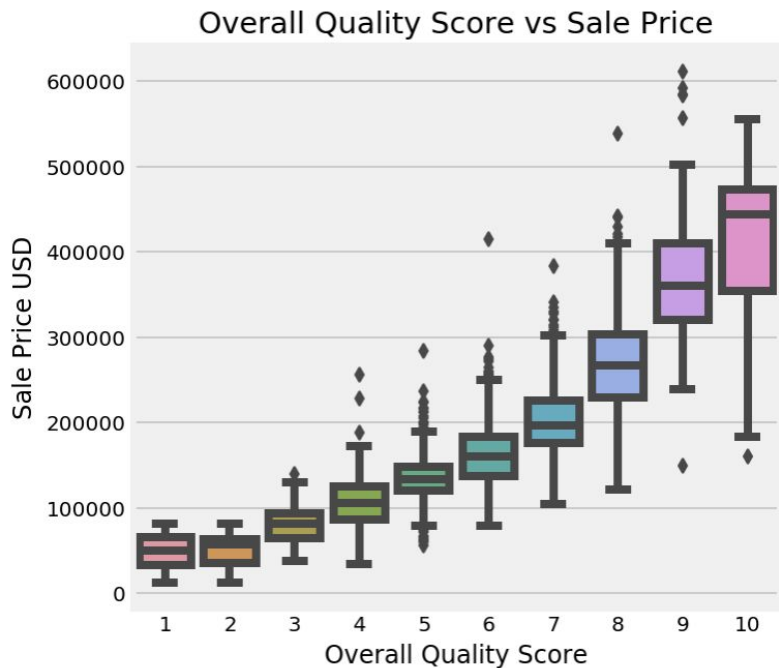
Data analysis revealed multicollinearity and skewness in the features

Sale Price Distribution
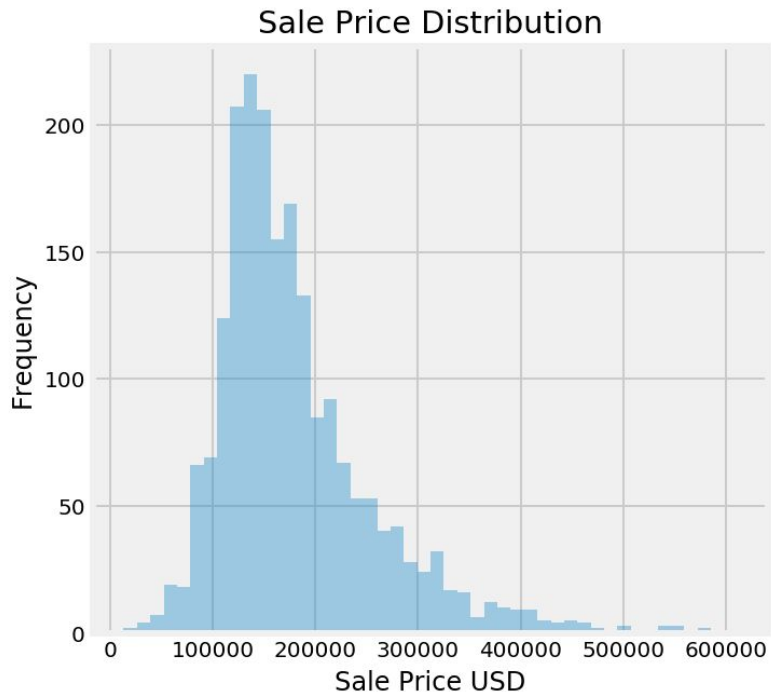
Reduction of multicollinearity:
- Identified highest correlated features by calculating Pearson's correlation
- Between highly correlated features (r > 0.6), dropped the feature with the lower correlation to Sale Price

Reduction of skew:
- Identified features with an absolute skew greater than 0.7
- Took the log of these skewed features to normalize their distributions

K-fold cross validation indicated LassoCV regularization has best baseline RMSE

Model selection process

1) Determined RMSE scores from 5 fold cross validation on unfitted base models

   a) Linear Regression

   b) RidgeCV

   c) LassoCV

   d) ElasticNetCV

2) LassoCV had lowest RSME followed closely by ElasticNetCV (high L1 ratio)

Model Evaluation Metric

$$\text{RMSE} = \sqrt{\sum \frac{(\hat{y}_i - y_i)^2}{n}}$$

**Lasso regularization may be best since the features have a high degree of multicollinearity**

Final fitted Lasso model had a RMSE less than $19,000 on test cross validation

Actual vs Predicted Sale Price

Key Takeaways

- Model performs best on cheaper homes and becomes less accurate on expensive ones

- Lower representation of high end homes in the training data likely account for this difference

- Model can be used to predict future home sale prices in Ames, IA given certain property features

Among top predictors of Sale Price are Total SF, Overall Quality, Lot Area, and Overall Condition

Lasso Model Top 10 Predictors of Sale Price
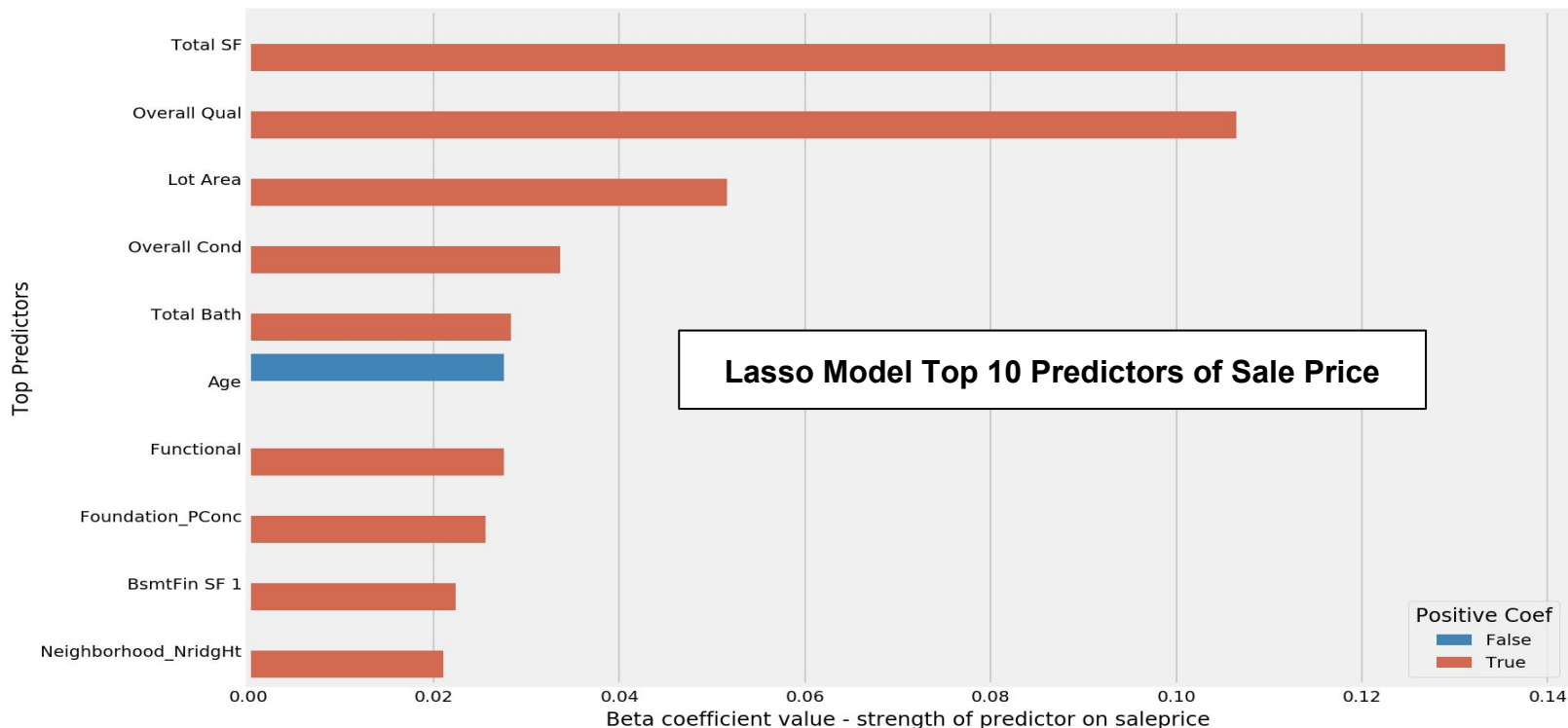
Model can be used to predict future home sale prices in a growing Ames, IA

Top Indicators of Home Price
- Total square footage
- Overall quality
- Lot area
- Overall condition
- Age
- Functionality rating
- Poured concrete foundation
- Finished basement square footage
- Northridge Heights neighborhood

Next steps

➢ Additional feature engineering to consolidate and reduce number of features

➢ Optimize hyperparameters (GridSearch)

➢ Explore different algorithms