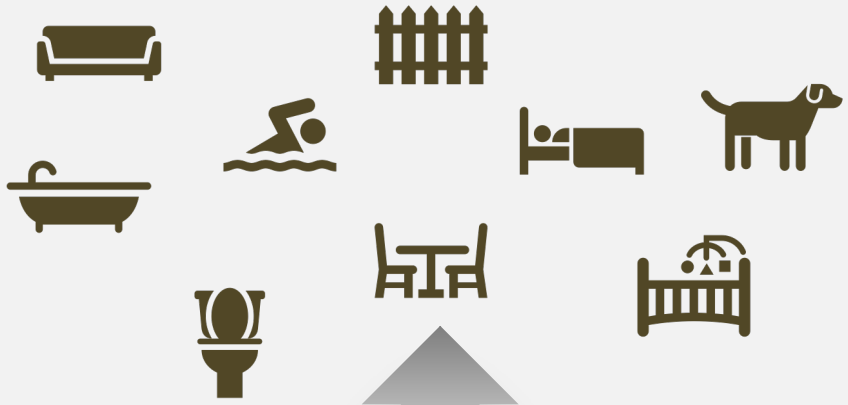
This Photo by Unknown Author is licensed under CC BY-SA

AMES HOUSING DATA

PREDICTING SALE PRICES USING LINEAR REGRESSION

BY JON GODIN



What Influences Home Purchase Prices?

How predictive modeling can help

Sellers

Are there things homeowners should improve to increase value? Which features may not be worth investing in?

Buyers

House hunters have differing wants and needs? Which ones matter more?

Pricing

We can help you predict what homes will sell for, helping reduce uncertainty for you and your customers

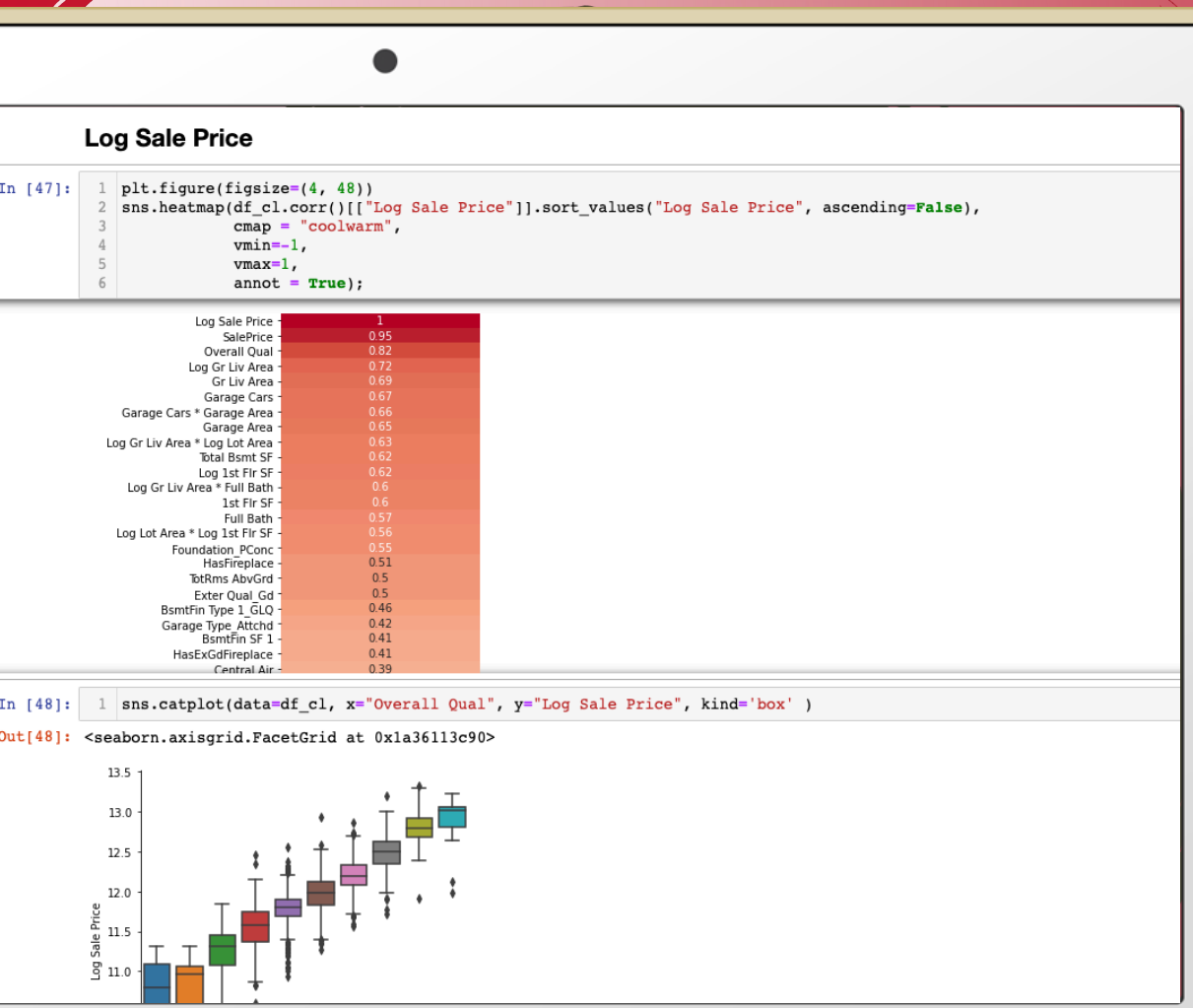
Data & Modeling Background

The Data

- In order to build our model, we used the Ames, Iowa Housing dataset, which contains roughly 80 different features of homes and the land they sit on
- Data includes home sales from **2006** to **2010**
- Two datasets were analyzed, a training data file of **2,051 home sales** was used to build the model, and a **separate file of 878 home sales was used for testing the model's performance**
- **Data Cleaning & Preprocessing:**
  - **Extensive data cleaning** was conducted to ensure that our data matched the original source data and was suitable for further analysis. Missing data was either appropriately recoded, or imputed using a predictive model (e.g., we used a model to fill in missing Lot Frontage data)
  - Categorical (non-numeric) features were converted to 0/1 **dummy variables**
  - **Feature engineering** included transforming variables in order to produce better modeling behavior of the data, creating **interaction effects**, and testing for **non-linearity** of relationships
  - Ultimately, a candidate set of **179 potential predictors** was created from the original set of 80 features

Modeling Log-Transformed Sale Price

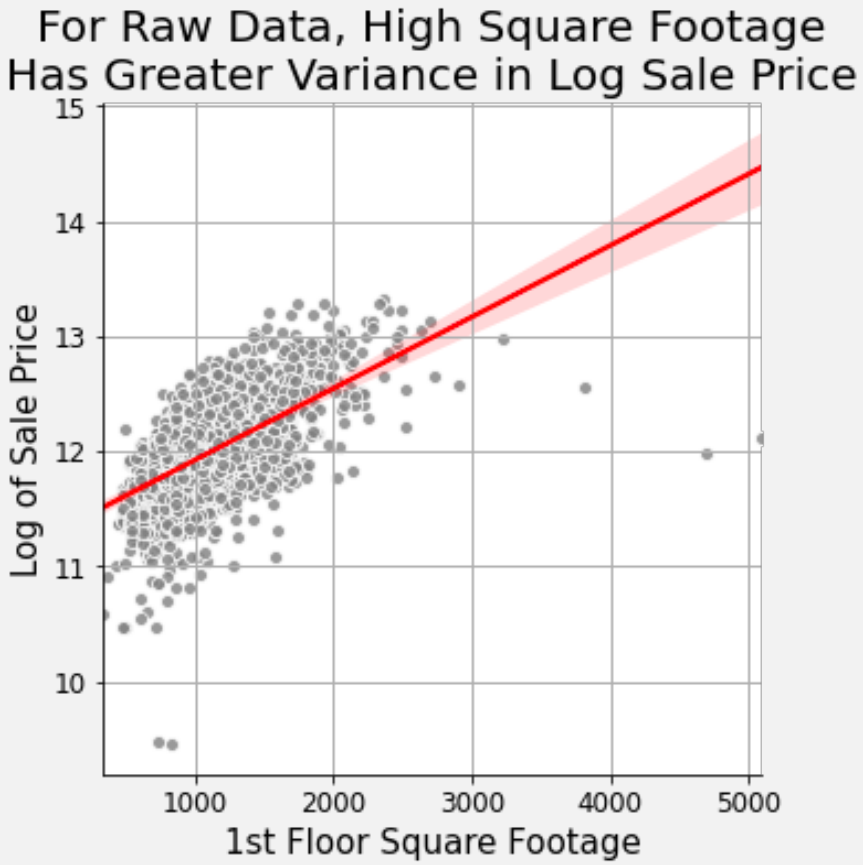
- Preliminary modeling guidance was provided via data visualization and exploratory analysis
- Modeling continued along four paths:
  - **"Traditional" regression modeling** with log(Sale Price) as the dependent variable and a mixture of log transformed and raw predictors
  - **Power-transformed regression modeling**, which is like the above except all data in the model gets transformed
  - **Ensemble regression modeling** which averages the results across multiple separate models to come to an overall prediction
  - **LASSO regression modeling resulted in the final, best predictions with the lowest amount of error**
    - Here all 179 features were included in the model, with the LASSO ultimately selecting which features contribute to improving out-of-sample prediction accuracy, setting non-impactful features to zero
    - The final model includes 107 of these features

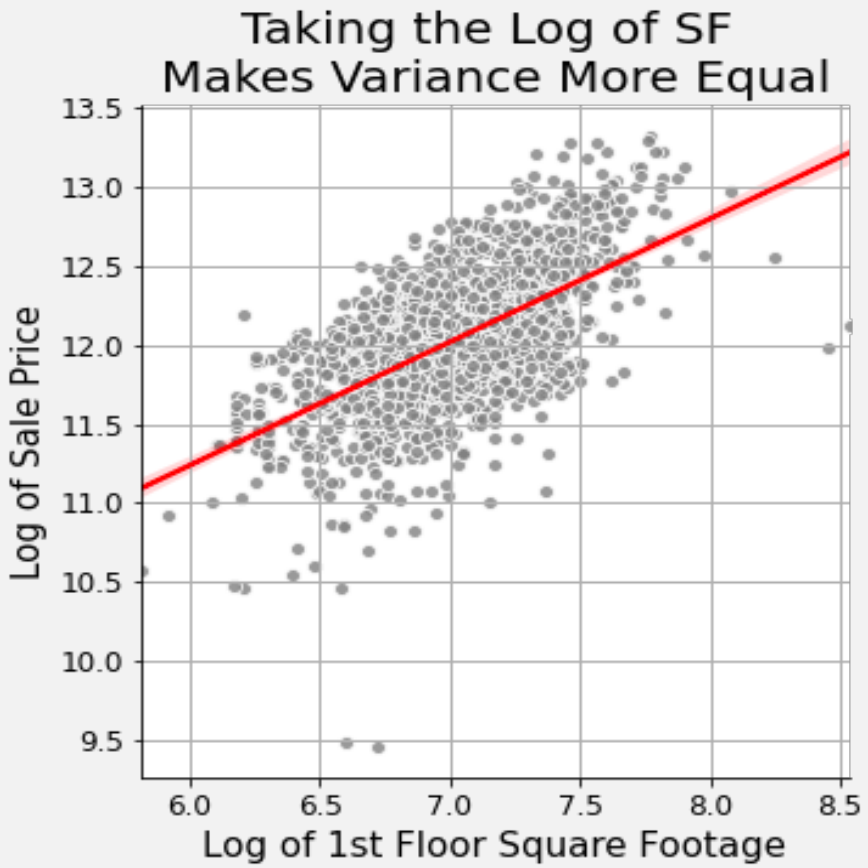


Exploratory Data Analysis

A few key highlights before we get to the final model

Quick Aside on Why We Log-Transform Variables





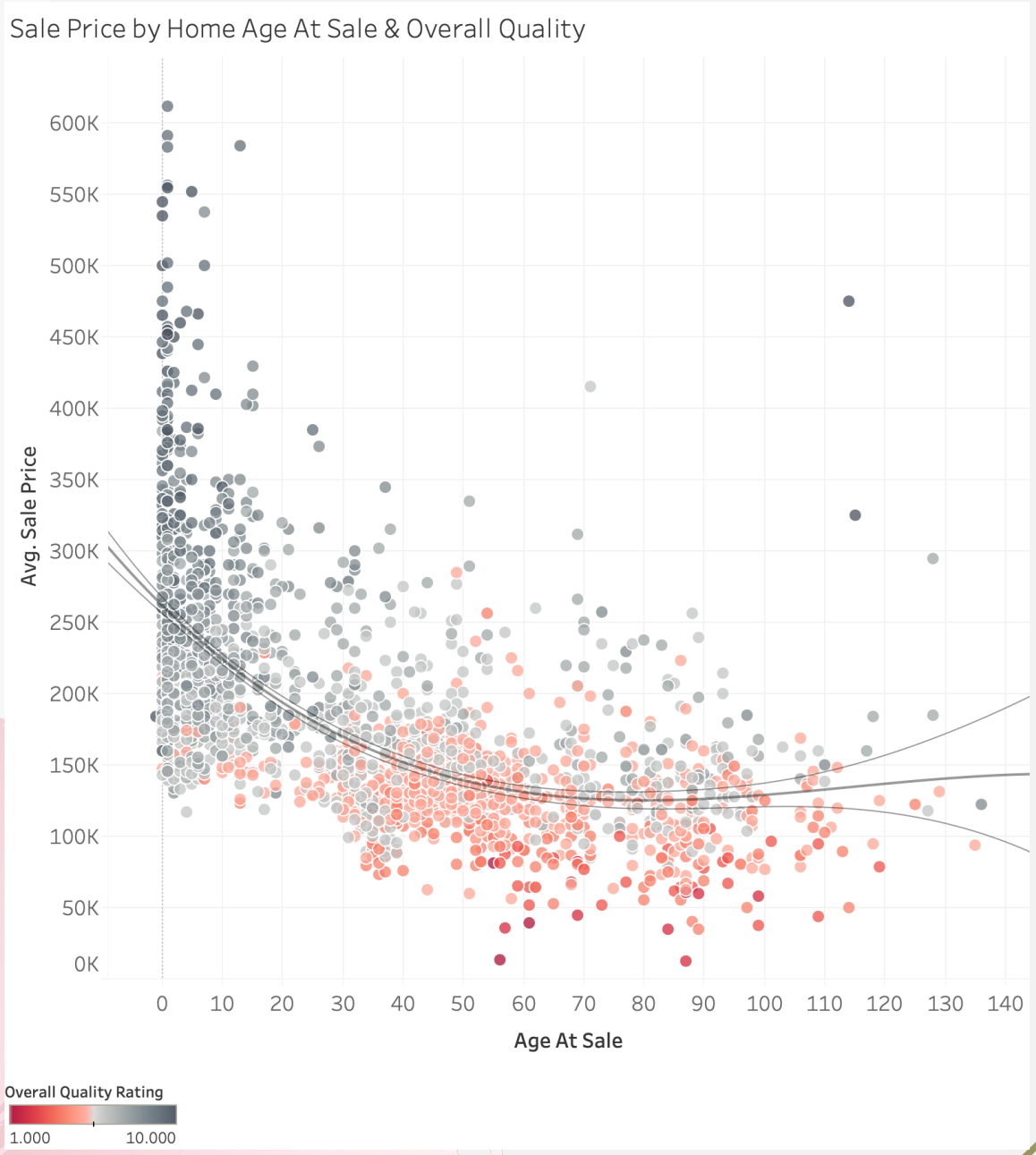
Note: a log-transformed value of 12 equates to a sale price of $162,755

Age At Sale & Overall Quality appear to be somewhat intertwined

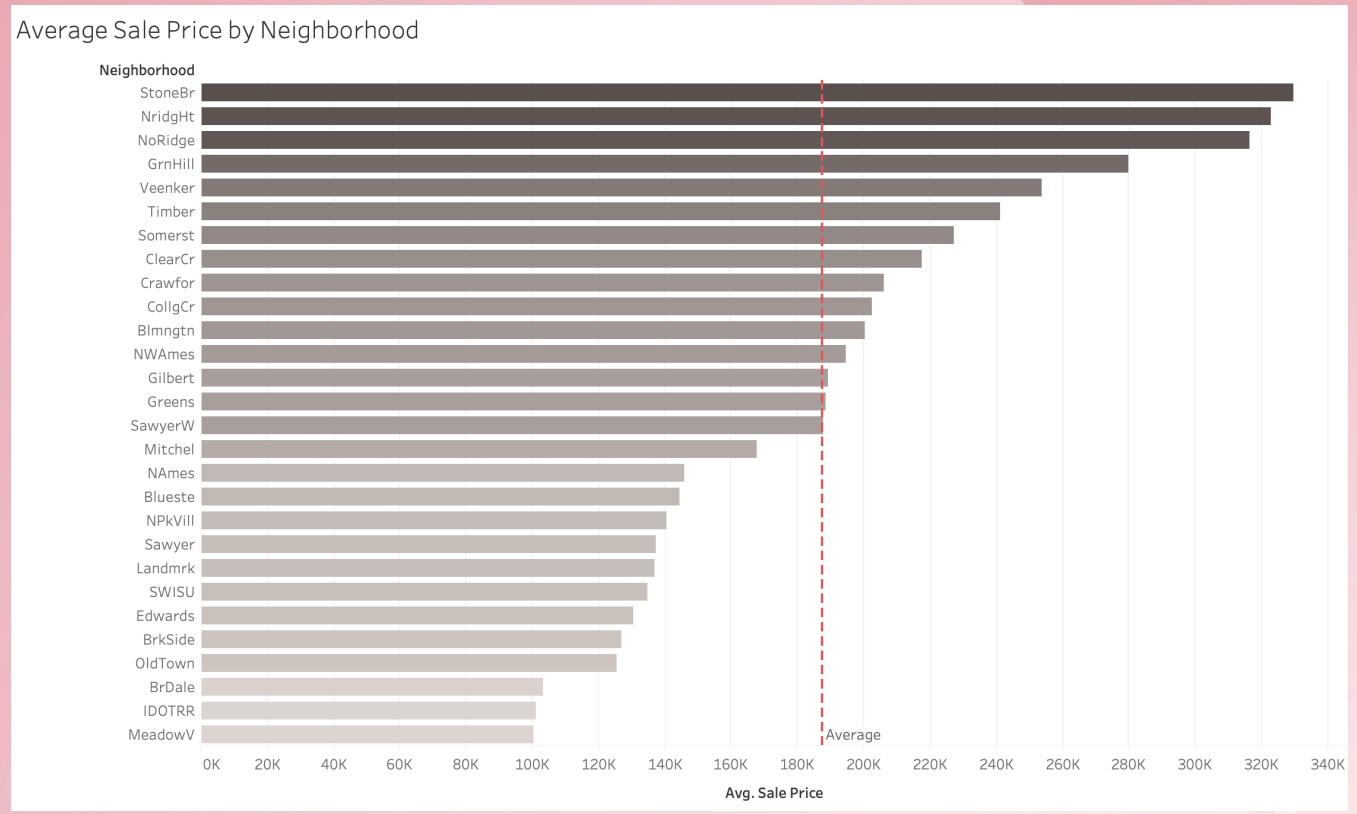
Despite this, the model does not include an explicit interaction effect between these variables

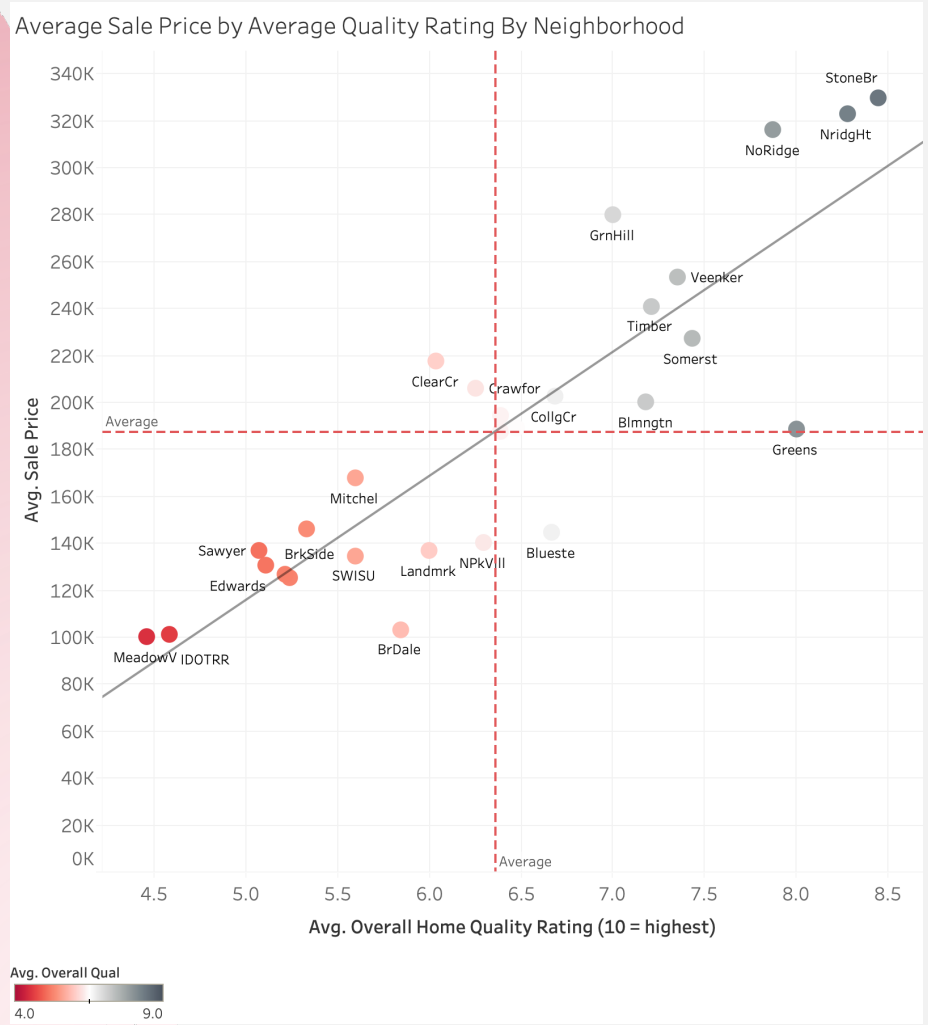
Inclusion of the interaction did not increase overall predictive accuracy

Very old homes in Ames have somewhat surprisingly-high quality ratings at time of sale, but maybe those few houses that lasted had better construction qualities to begin with

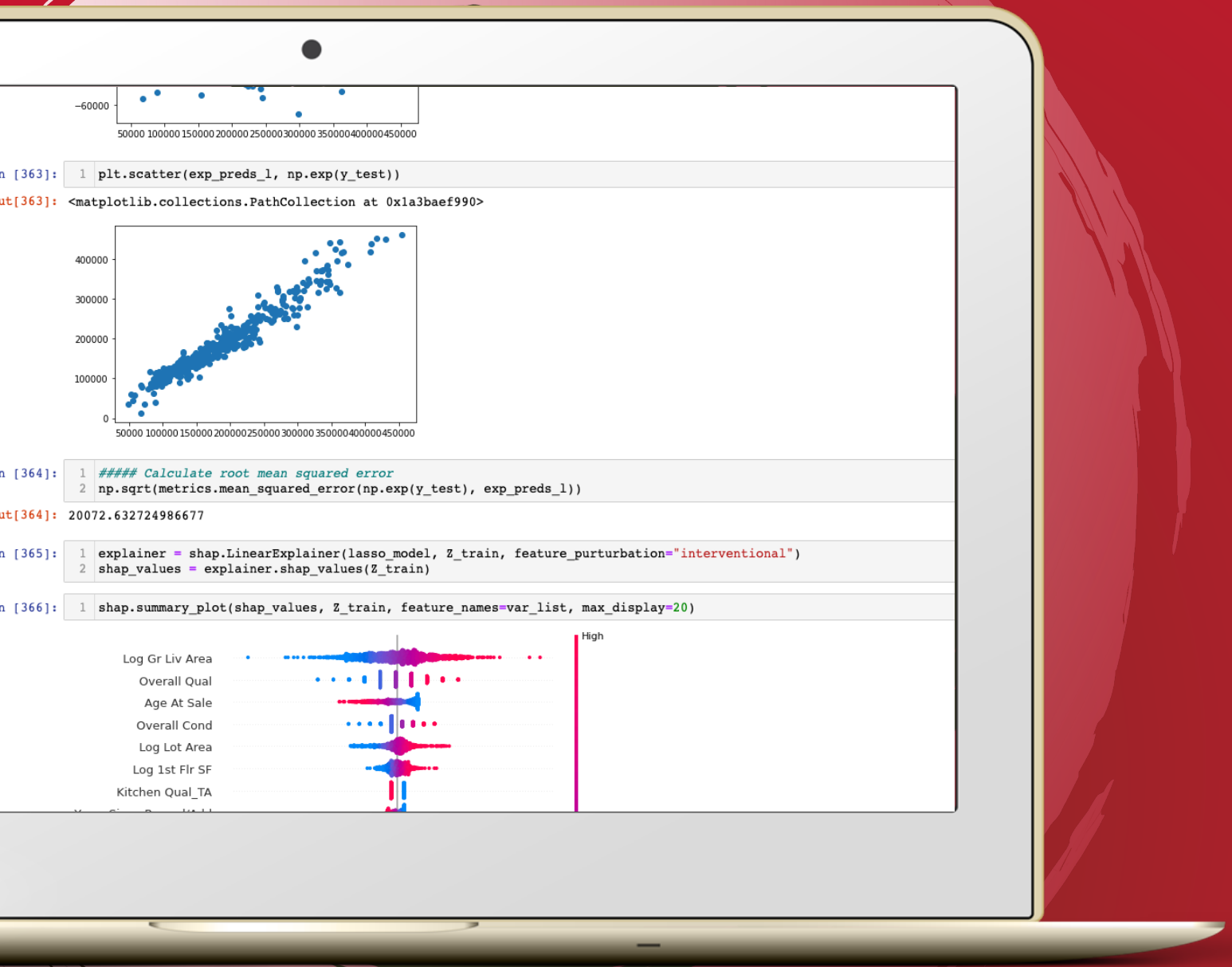


Neighborhood appears to impact average house prices





While we do include neighborhood in our model, the final effects are not as large as they appear here after we control for the other variables in the model

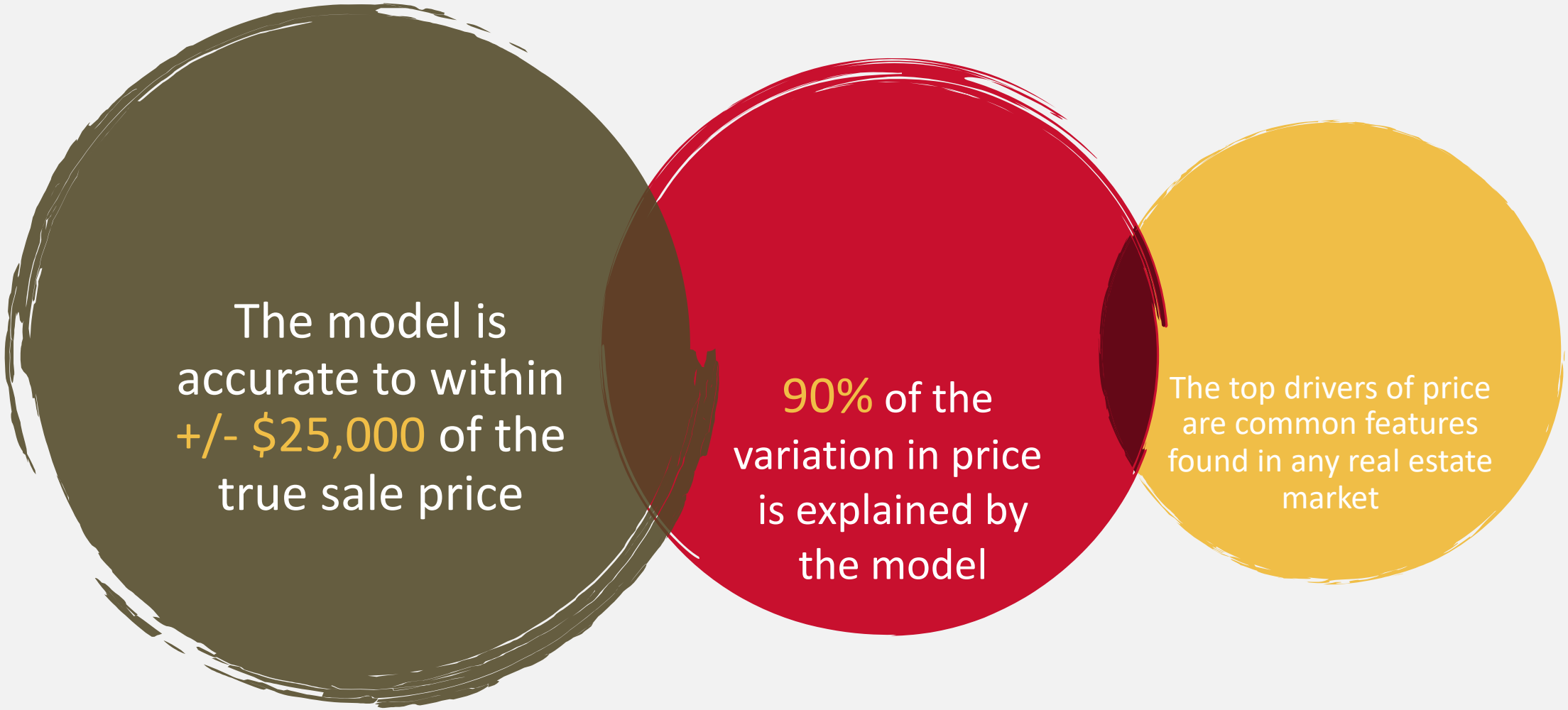


Model Results

How accurate is the model?

What drives sale price?

Key strengths of the model

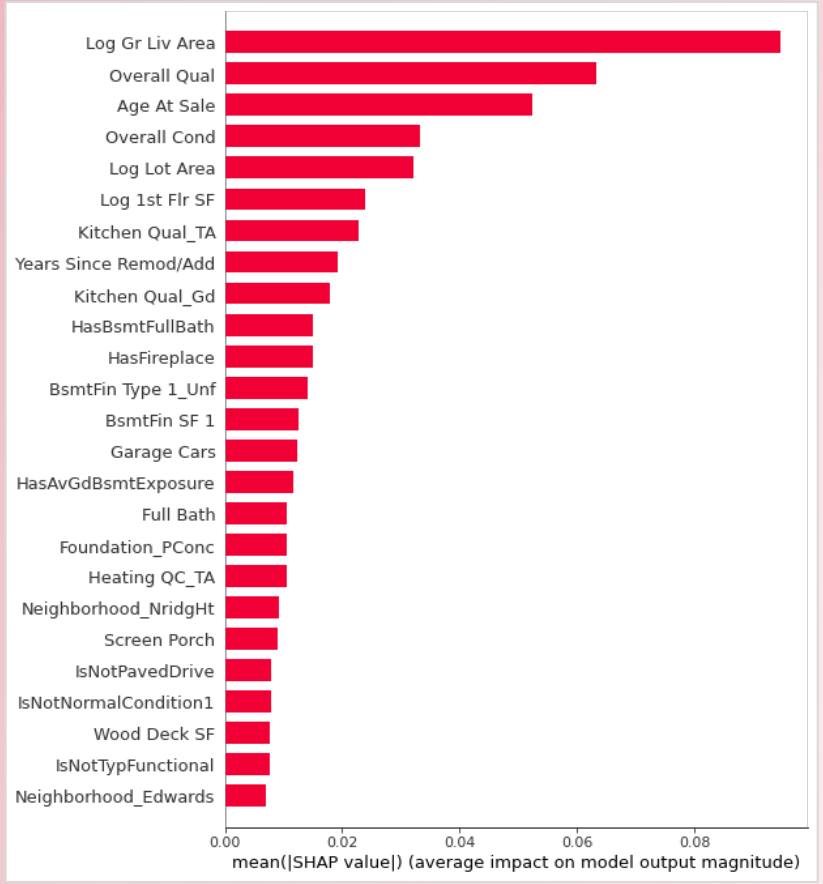


Gross Living Area, Overall Quality, & Age At Sale Drive Price the Most

Using a technique from Game Theory, we can assess the relative impact of each feature on price changes

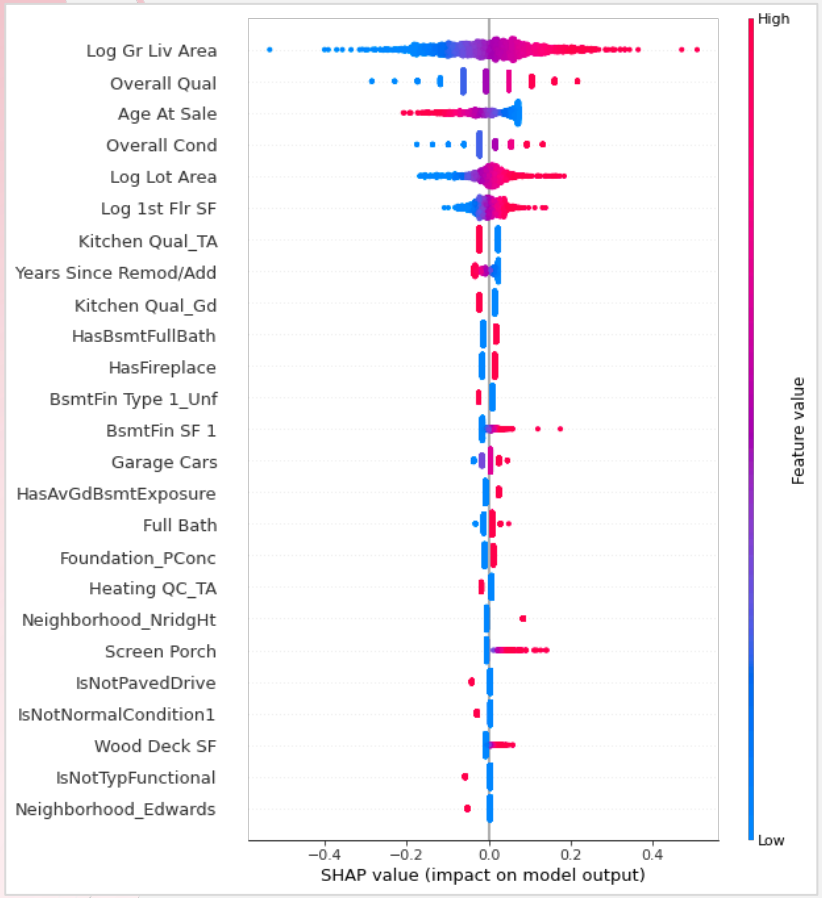
Top 25 Predictors

Absolute Relative Impact on Price



Top 25 Predictors

Distribution of Impact on Price



We can also value features based on the model

Three features are highlighted here, but we can translate any additional features you are interested in

Overall Quality Rating

A 1-point increase in quality rating equates to an 8% increase in average sale price, holding everything else in the model equal

Gross Living Area

A 10% increase in gross living area equates to a 1.1% increase in average sale price, holding everything else in the model equal

Age At Sale

For each additional year in a home's age at time of sale, the average sale price decreases by 6.3%, holding everything else in the model equal

Weaknesses of the model

Many of these could be tweaked for other markets

Complexity

Because there are a lot of features in the model, it can require more time to fully understand, and some of the effects may appear small even though they are reliably predictive

Neighborhoods

These increase the accuracy of the model, but of course are only relevant for the Ames market

New neighborhood dummy variables can be inserted instead for a new market

Local Variation

Certain house types or zoning classifications may be more common in Ames than elsewhere, and vice-versa, so some additional adjustments of the model may be needed for a new market

Data Is Now Old

The Ames Housing dataset is feature rich, but the most recent sale date was 2010, so a decade has now passed

Adding new sources of data to the model with greater recency would add further improvement

Conclusions & Recommendations

- The LASSO regression model provides the dual benefit of high predictive accuracy and inclusion of enough features to understand what drives value
- Key drivers of sale price include:
  - Gross living area and 1st floor square footage
  - Overall quality and condition ratings of the house
  - Age at sale and years since the last remodel/addition
  - Lot area
- Additional features with high value include:
  - Excellent kitchen quality
  - Finished basement with average or better exposure
  - Full bath in the basement
  - Fireplace
  - At least a 2-car garage
  - 2 or more full bathrooms in the house
  - A screened porch, with larger sizes being more valued
  - A paved driveway
- Buyers should know they will need to pay more for larger homes, larger lots, newer and better-quality homes, all else being equal
- Sellers should focus on maintaining the house, and making improvements related to the key features listed above

THANK YOU