Hunting for Features that Matter

Bibor Szabo

Problem Statement

1. Which features of a home affect the sale price the most?

2. Which machine learning model predicts sale price better?

Data

Ames Housing Data

- Collected in Ames, Iowa
- By the Ames Assessor’s Office
- Individual Residential Properties
- Sold between 2006 - 2010

- Number of observations: 2051
- Number of variables: 81

http://jse.amstat.org/v19n3/decock/DataDocumentation.txt

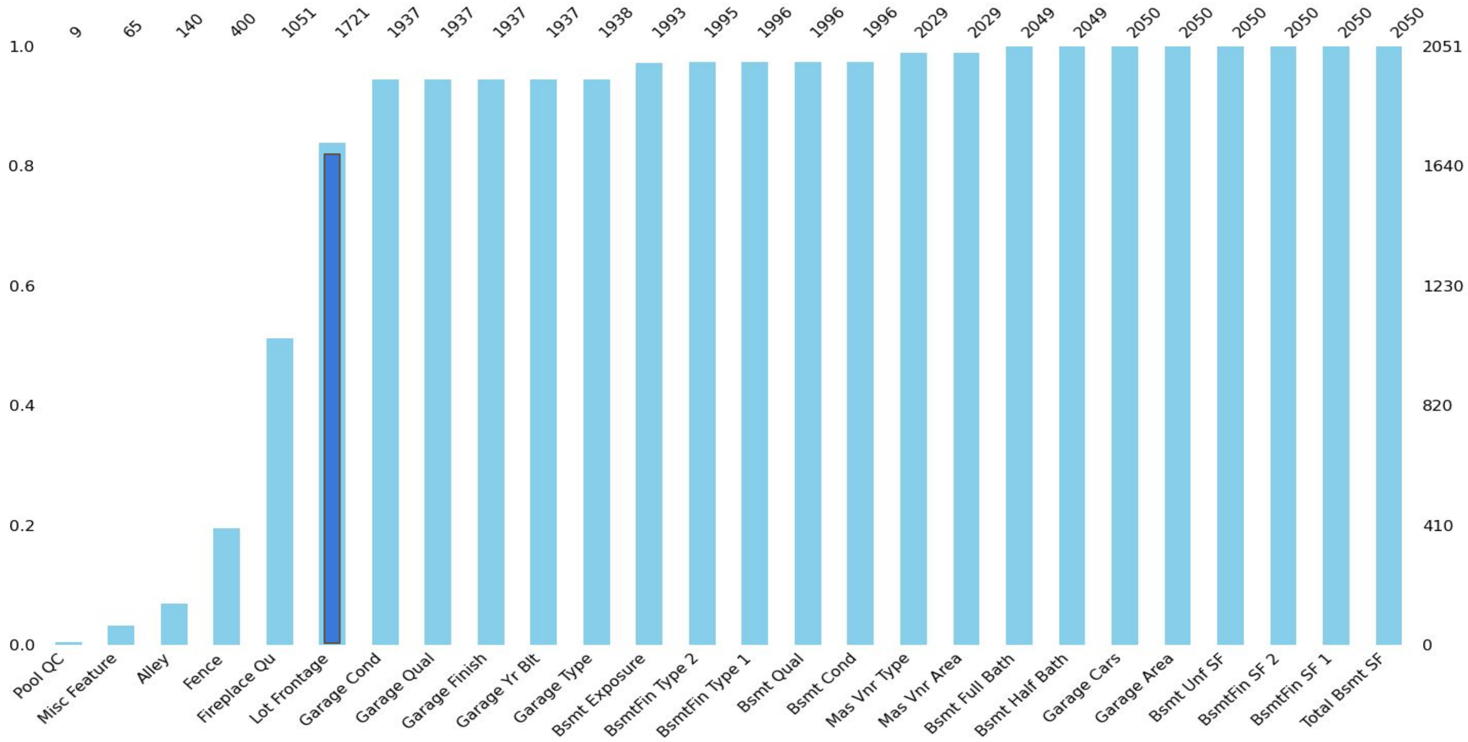
Features	Examples
- 23 nominal	Zoning Classification, Lot Shape, Neighborhood, etc.
- 23 ordinal	Land Slope, Overall Quality, Overall Condition, etc.
- 14 discrete	Year Built, Month and Year Sold, Number of Bedrooms, etc.
- 20 continuous	Sales Price, Areas of Different Features
- dependent variable continuous	Sales Price

Challenges

- Number of Features
- Number of Categorical Variables
- Missing Values

- Outliers

Missing Data



Feature Engineering I.

Recoded Variables:

- Year Built → Age when sold
- Area → Total Squarefeet

Dummies:

- Nominal variables → Binary by category
  (Zoning, Alley Type, Utilities)
- Ordinal variables → Binary by category
  (Year Sold)
- Continuous variables → Has Porch, Has Pool
  (Porch Area, Pool Area)

Feature Engineering II.

Polynomial Features:

- Interaction variables
- Squared variables
- Excluded bias

Standard Scaler

Models

Unregularized:

Linear Regression

Regularized:

Ridge Regression

LASSO Regression

Model Evaluation I.

Linear Regression

- Manually Selected Features
- Correlation Matrix (10 best)

- Train R-Squared: 78.2%
- Test R-Squared: 84%
- Cross-Val-Score: 76%

Linear Regression - Automated

- SelectKBest Features (45 best)

- Train R-Squared: 87.9%
- Test R-Squared: 87.4%
- Cross-Val-Score: -2.3x10^23
    - 89.7%
    - 85.6%
    - 84.4%
    - 44.1%
    - -1.14x10^24

Model Evaluation II.

Ridge Regression (RidgeCV)

- All original features (except Id and PID)
- All polynomial features

- Train R-Squared: 98.5%
- Test R-Squared: 90.4%

- Overfit

LASSO Regression (LassoCV)

- All original features (except Id and PID)
- All polynomial features

- Train R-Squared: 99.9%
- Test R-Squared: 84.4%

- Overfit

Conclusion

- Models: more complex ≠ better

- Feature Selection: machine learning techniques ≠ better prediction

Recommendations

- Preference of Linear Regression Model

- Features:
  - Most effect: overall quality and total area of home
  - Machine aided - intuitive feature selection

Features that Matter

Condition
Size

Resources

http://jse.amstat.org/v19n3/decock/DataDocumentation.txt