Predicting Housing Prices

Will Herzog

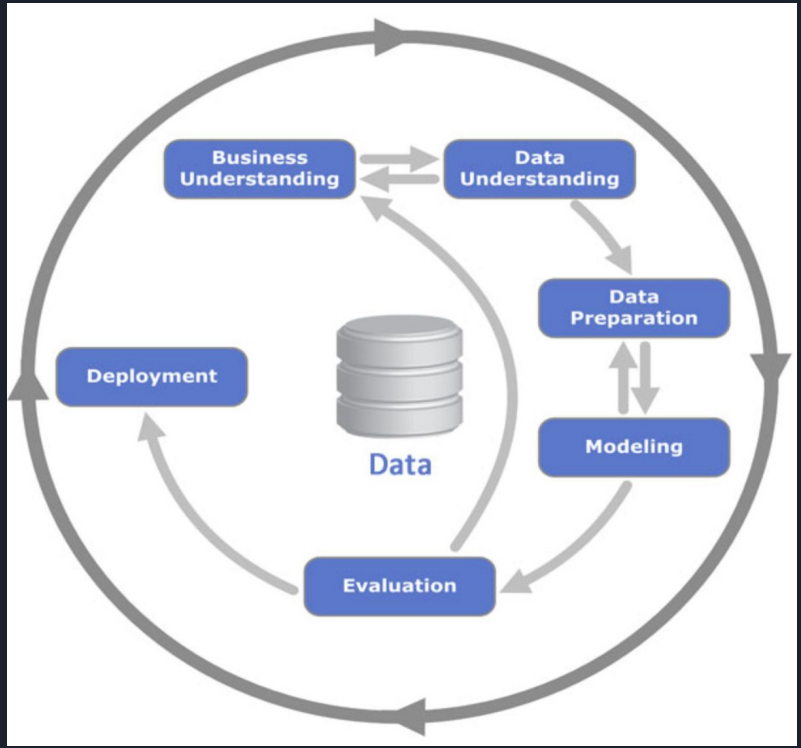
Objective - Determining the expected house price

- Developing a model
  - Gathering data
  - Creating and refining a model
  - Drawing conclusions from the model
- What components of a house influence the price the most?

Overview - Process for creating the model

1. Data Understanding

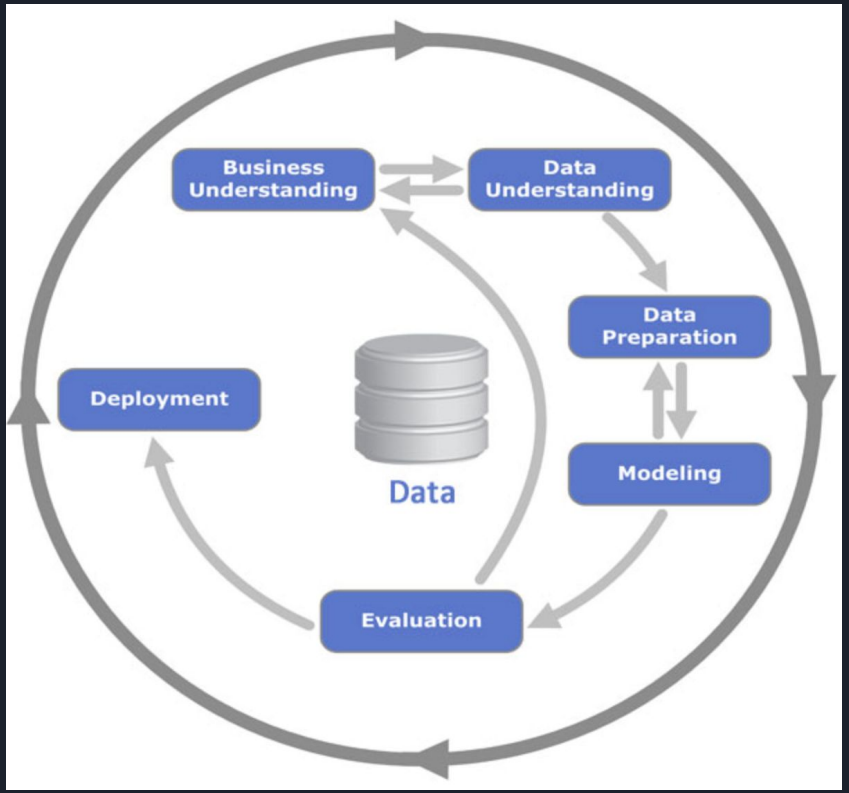
- Kaggle DataSet
  - https://www.kaggle.com/c/house-prices-advanced-regression-techniques/
- Data Shape
  - 1460 houses (rows)
  - Initially 80 features
    - ~ 2/3 categorical



Data Preparation

2. Data Preparation

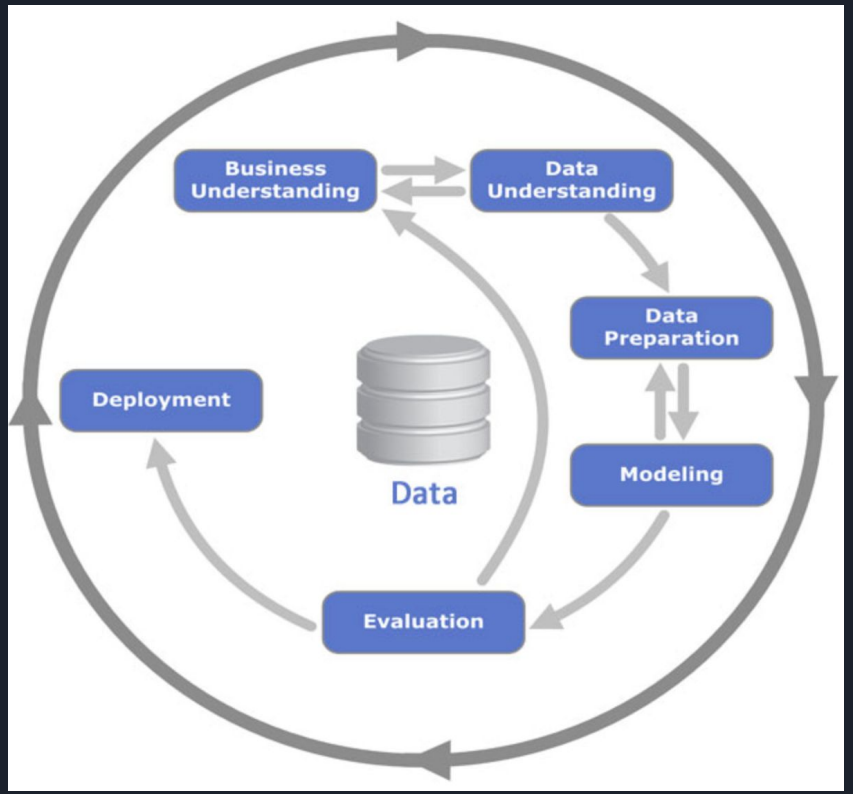
- Data cleaning
- Numerical Features
  - Feature scaling
  - Transformation
- Categorical Features
  - Getting dummy variables
    - Total features 213



Model Description

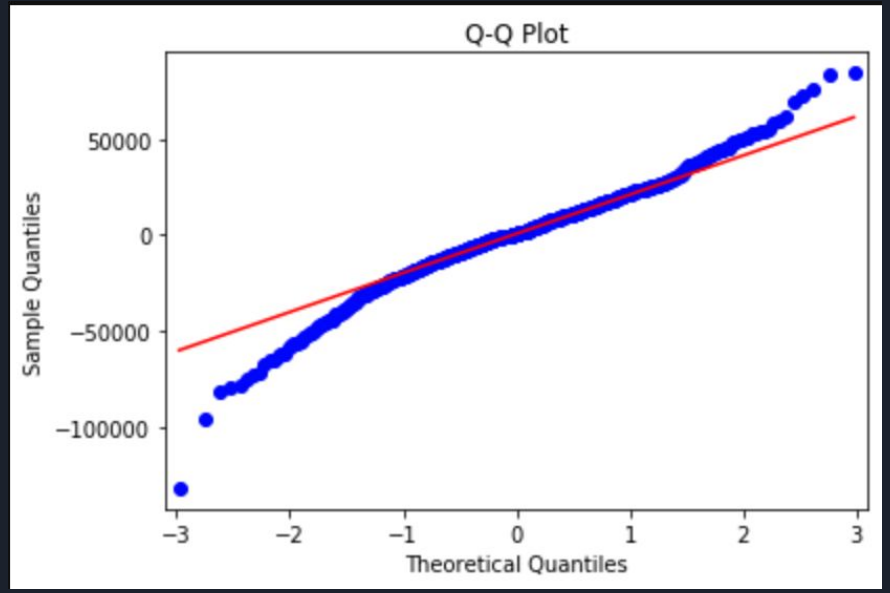
3. Modeling:

- Linear Regression
  - R^2: .65 (initial)
  - R^2: .817
    - Poor adj R^2
- Lasso Regression
  - Eliminates low importance features
    - 213 -> 182



Checking Linear Regression Assumptions

- Linearity
- Normal distribution of residuals
- Homoscedastic
  - Residuals have constant variance around zero
- No correlation between features
- No autocorrelation
  - Durbin - Watson: 2.1



Model Results

4. Evaluation - Using Lasso Regression

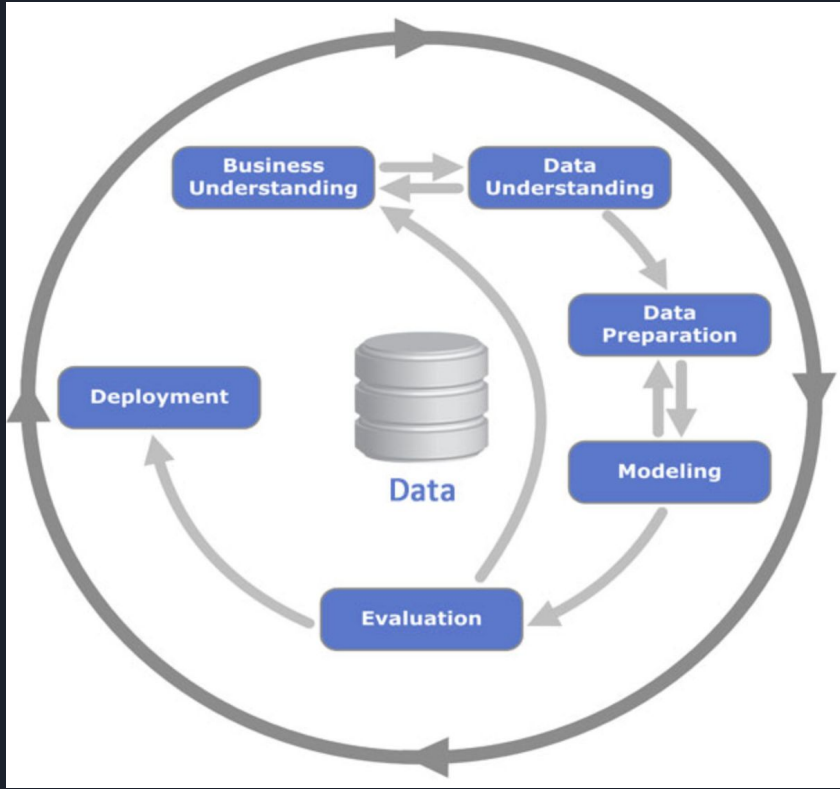
R^2: .902
Adjust R^2: .7651

Top 5 Features:

1. 3 Season Porch sqft
2. Screen Porch sqft
3. Basement sqft
4. First Floor sqft
5. Above ground living area sqft

Top 5 Negative Features:

1. Lot Area
2. Garage Area sqft
3. Enclosed Porch sqft
4. Year Built
5. Low Quality sqft



Recommendations

- Incorporate more data
  - Generalize the model further
    - Improve future results
- Most significant factors in the model

Questions?