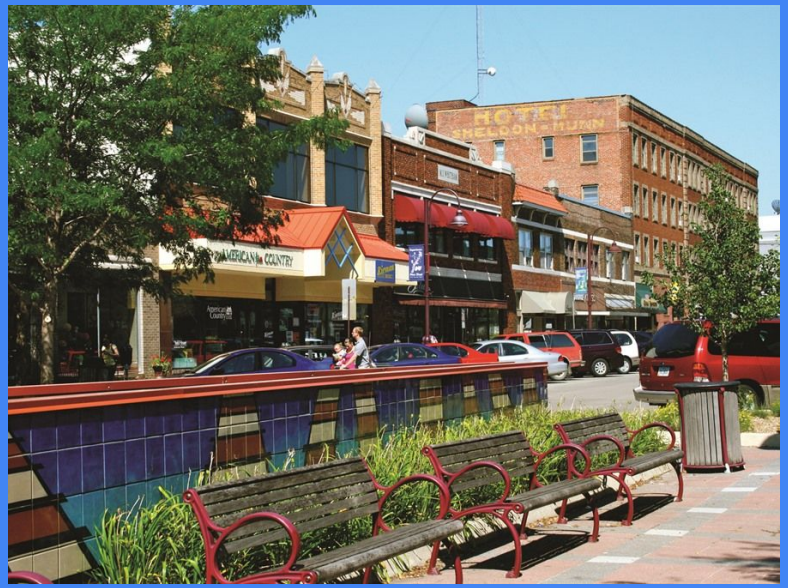
Predicting Home Sale Prices in Ames, IA

Maria Yarolin



Background

Goal: Employ predictive modeling (i.e. using existing data to develop a model that predicts an outcome) to predict housing sale prices

Target audience: Real estate professionals with a general knowledge of the analytics relevant to their industry, but whom are not data scientists

Data source: *Ames Housing Dataset* - contains over 70 variables related to homes sold in Ames, IA from 2006-2010; the dataset to build the model includes 2051 cases.

Techniques Used: Preparing the Data

Pre-processing and examining the data:

- **Fill in missing data** on potentially useful variables
  - Minimal alteration, limited to only a few missing data points
- Determine **correlation** (i.e. the strength of relationship between two variables) of Sale Price and the other variables
  - Variables with a moderate-to-high correlation with Sale Price were selected for further consideration
- **Rescale** the data to mitigate the effects of variables with different units of measurement

Techniques Used: Predictions

Modeling:

- **SelectKBest** - determines what features to include in a model by removing the lower-performing ones and retaining only the best features
  - Features selected for the model: Overall quality, Ground floor living area sq.ft, Garage area, # of cars in garage, Total Basement sq.ft, 1st Floor sq.ft, Total rooms above ground, Year built, Year remodeled/added

- **Multiple Linear Regression** - models the relationship between two or more predictor variables and a response variable (Sale Price) by fitting a linear equation to observed data

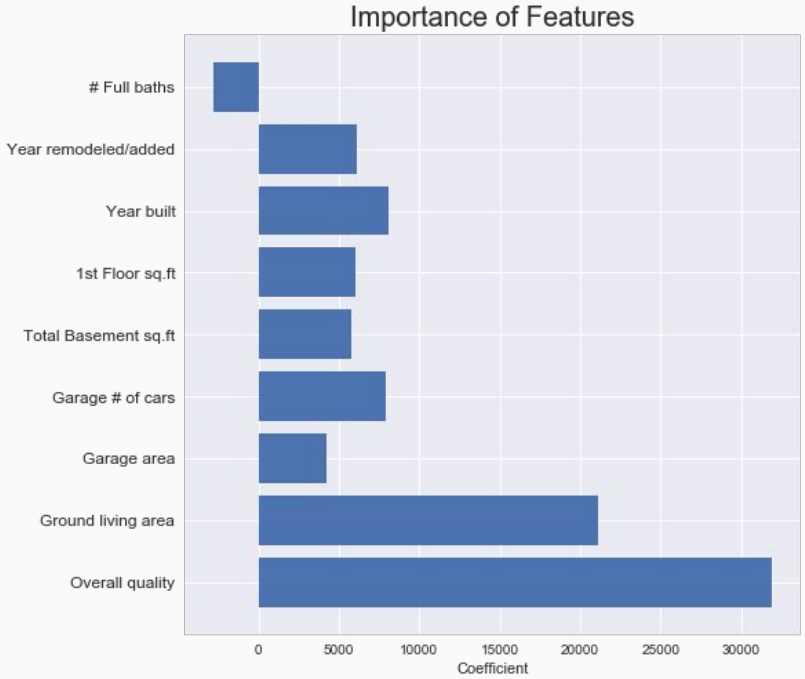
Model Performance: Price Predictions

The closer the points are to the line, the better the fit.

The model is good at predicting sale prices at the low-to-moderate range, but is less effective for predicting prices at the higher end of the range.

The R^2 score is 0.815481856581, meaning that this model explains 81% of the variation in housing prices. The remaining 19% of variation is attributed to other factors not addressed by this model.

Model Performance: Feature Importance



By far, **overall material and finish quality** is the most important feature in determining housing prices, followed by the **square footage of living area above grade (ground)**.