House price prediction algorithm

Alix Benoit

6/16/2020

Introduction

(describes the dataset and variables, and summarizes the goal of the project and key steps that were performed.)

This project uses the Ames Housing Dataset. This dataset includes information about 2930 house sales from 2006 to 2010 in Ames, Iowa, and contains 79 different explanatory variables. It was put together by Dean De Cock from Truman State University; more information about it can be found here.

I acquired this dataset from an ongoing "getting started" kaggle machine learning competion: House Prices: Advanced Regression Techniques

For the purposes of writing this report and calculating my final RMSE, I also downloaded the actual house prices for all entries from the original dataset (Note that this would not be possible if this were not a "getting started competition).

The following libraries were used:
- Tidyverse
- Caret
- Kknn
- Rborist

Of the 79 explanatory variables, 43 were identified as categorical and 36 numeric.

The goal of this project is to accurately be able to predict house prices in Ames, Iowa based on a variety of different factors. It is also to learn more about feature engineering/machine learning and to apply skills gained in Harvardx's proffessional certificate program in data science.

The root mean squared error of the log predictions was chosen as the measure of accuracy of the models; the log is taken so that errors on expensive houses will be weighted the same as errors on cheap houses.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln\hat{Y}_i - \ln Y_i)^2}$$

In short, the final model was obtained by:
- Combining explanatory variables from the given train and test set in order to impute NAs
- Splitting full set of explanatory variables back into a train and test set
- Obtaining the best measure of center for ratings in the train set.
- Fitting a weighted knn model to the numerical variables of the train set, predicting centered rating
- Fitting a random forest model to the categorical vriables of the train set, predicting centered rating, minus predictions from the knn model.

Method/Analysis

Results

Conclusion