Predicting Sale Price of Homes in Ames, Iowa

Kyle Jones (gtid3: 449, email: [kjones374@gatech.edu](mailto:kjones374@gatech.edu))

*November 28th, 2021*

*Data Mining and Statistical Learning (ISYE-7406), Fall 2021*



# Abstract

# Introduction

Accurate projections of a home’s sale price are relevant to individuals interested in buying or selling a home and to those aiding in the process (e.g., real estate agents, brokers). Some simple features related to a home’s sale price that are available in almost any housing dataset include square footage, location, number of bedrooms, number of bathrooms, etc. These variables alone often explain much of the variation in sale price, especially if you control for the year the house sold.1 However, there are many other attributes of a home that may be relevant to the sale price, that often aren’t available in housing datasets (e.g., proximity to park, foreclosure or normal sale, zoning classification of sale).

The Ames housing dataset contains 80 variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) related to the quality and quantity of physical attributes of 2,919 home in Ames, Iowa that sold between 2006-2010.2 Most of the categorical variables contain many unique values (e.g., 25 unique neighborhoods), which when combined with the possibility of feature engineering (e.g., interactions between variables, transformations) makes it possible to create hundreds (maybe even thousands) of predictor variables. As a result, the Ames housing dataset was the perfect dataset to explore less commonly available predictors and assess their relevance to sale price predictions.

The primary goal of this project was to construct the best possible model to predict sale price using techniques taught in ISYE-7406. Due to time restrictions, I was not able to explore all possible predictor variables, so the models created used information from only 25 of the 80 variables available. Given unlimited time, I would have explored many more variables and most likely developed better models. This report will walk you through the methods I used to predict sale price, my strategies to create the best model, estimated sale price errors on unseen data, and actual errors on an unlabeled test set. Additionally, the relative importance of home attributes that I examined will be discussed.

# Data Source

This was a Kaggle competition dataset:

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

A Kaggle competition dataset means that sale prices were only available for the training set. After model development, the best model was used to make predictions on the unlabeled test set. This allowed for error calculations on unseen data.

# Proposed Methodology

# Analysis and Results

# Conclusions

# Appendix

# References