Project 1: Predict the Housing Prices in Ames

Contents

Ames Housing Data

Dataset Overview

The dataset for this project is available for download at proj1.zip. Once you unzip the file, you'll find **ten** folders. Within each folder, there are **three** files:

- `train.csv`: This file represents the training dataset and contains 2051 houses across 83 columns.
  - The first column is "PID", the Parcel identification number;
  - The last column is the response variable, Sale_Price;
  - The remaining 81 columns are explanatory variables describing (almost) every aspect of residential homes.
- `test.csv`: This is the test dataset containing feature vectors for 879 houses.
- `test_y.csv`: Complementary to the `test.csv`, this file contains the response column (Sale Price) alongside the "PID" column for the test set.

Note: Students are required to handle each folder independently. For example, you should train models using `train.csv` and evaluate them using `test.csv` from fold1, and repeat this process through fold10, treating each fold as a distinct dataset and not combining the data across folds.

Dataset Origin

The training and test splits are derived from the Ames Housing data. For more background on this dataset, you can refer to:

- De Cock, D. (2011). "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project," Journal of Statistics Education, Volume 19, Number 3. [PDF]
- Check variable description [Here]
- This dataset also features in a Kaggle competition (https://www.kaggle.com/c/house-prices-advanced-regression-techniques). Our dataset, however, has two additional explanatory variables: "Longitude" and "Latitude". Exploring the Kaggle competition can offer insights on data analysis approaches and sample codes.

Project Objective

Your task is to predict the price of homes, but importantly, **in log scale**. You need to build **TWO** prediction models selected from the following two categories:

- one based on linear regression models with Lasso or Ridge or Elasticnet penalty;
- one based on tree models, such as randomForest or boosting tree.

Note:

- The features selected for the two models can differ.
- `PID` cannot be used as a feature. `PID` is a unique identifier for parcels of land or properties assigned by the county. It's more like an index and has no logical connection to housing price determinants.
- Please refer to Campuswire to identify the packages that are permissible for use in this project.
