Kaggle: House Prices: Advanced Regression Techniques

Team name: abmasadullah

Team Members: A. B. M. Asadullah; ID: 1640CSE00523,
Md. Al-Amin; ID: 1640CSE00504,
Makhjum Deen; ID: 1640CSE00500

Links: Kaggle: here; GitHub: here

Project Goal

It is your job to predict the sales price for each house. For each Id in the test set, you must predict the value of the SalePrice variable.

Problem Statement

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

Data Preprocessing

This project collects 1,457 houses from three prominent different housing websites. The reason is to make sure that the scraped housing data is as accurate as possible. Since all three websites get data from listing services or other thirdparty companies, inconsistency and mistakes in data are unavoidable.

All three websites record that the house was last sold in October, 2017. However, the sold price of the houses is inconsistent. On Zillow and Redfin, the house's sold price is recorded as $325,000 while the price is $233,427 on Trulia. In order to handle this consistency, I simply pick the number that appears most often (which, in this case, would be $325,000). If the three websites record three different sold prices, then I take an average of the three and mark the house for a later check-up. In case that the sold prices vary in a wide range across the three websites (e.g. $1 in Zillow, $16,000 in Trulia and $58,000 in Redfin), the observation will be drop. Besides sold price, other housing characteristics are also subject to the same comparison algorithm.

Besides inconsistency in data values across three websites, there is also inconsistency in data units. For example, size of a house can be recorded in either square feet or acres. Therefore, an extra conversion step has to be taken in order to uniform data units. All data processing steps are done in Excel's Visual Basic for Applications (VBA).

Data Description

In this dataset, there are 35 housing attributes, including internal attributes and external attributes. Internal housing attributes, such as number of bedrooms and number of bathrooms, are intrinsic variables to the houses. On the other hand, external housing attributes, such as the walkability of the neighborhood and public schools' scores, are variables that are not builtin with the houses.

For example, for the non-numeric attribute Sold Month, its dummy variables are the twelve months of the year. If a house is sold in January, then the variable January would take a value of 1, and 0 otherwise Feature Engineering (feature subset selection, creation, dimensionality reduction) Modeling Methods (cross-validation, ensemble)

Results & Discussion

For short time and our short coming we have submitted only 1 (one) time and our position is 2356.