101C Project Report

Avash Monajemi, Yuqi Zhu, Jaehyeong Lee, Iris Liu

December 5, 2018

1 Project Goal

The various quality and quantity of many physical attributes of a property provide new oppurtunities for prediciting the affordibility of a home in Iowa. There are 79 variables specifically which are directly related to property sale, and hence the affordability of a home. The challenge and goal is to clean the data in such a way that most robust classification techniques will yeild accurate predictions that result in the lowest misclassification rates.

1.1 Problem Statement

It is your job to predict the affordability for each house. For each Id in the test set, you must predict the value of the affordability variable. (Affordable vs. Unaffordable)

2 Data Preprocessing Methods

Examining the dataset briefly gives preliminary insight for how the training and testing data should be manipulated which are : (1) Certain characteristics of a home have multiple variables linked to them, (2) Data collection from select predictors resulted in a number of NA results that ought to be corrected for, and (3) Existence of both numeric and character variables and how they may conflict during future data modeling procedures unless they are restructured or simplified. The following outlines the most common methods that are used to manipulate this data along with several examples and motivations for doing so.

Deletion of a Predictor Entirely

1. **Evidence claimed the predictor is not linked with affordibilitty**

- *Example* : Removal of the LotArea predictor due to evidence claiming that you cannot neccacarily determine the price of a home based on lot area.

2. **Existence of other predictors already explain the predictor**

- *Example* : BsmtFinType1 and BsmtFinType2 describe the rating of the basement finished area. After we take into account the existence of the predictor BsmtCond which evaluates the general condition of the basement, these variables now become obsolete by themselves as the general condition of the basement explains it enough.

3. **Predictor result is embedded in creation of new predictors**

- *Example* : FullBath and HalfBath variables are accounted for (totaled) in a new predictor which is total # of baths.

4. **Single level dominates the predictor (or too many NA's)**

- *Example* : The majority of homes in the data do contain central air conditioning (CentralAir : Yes or No), so analysis would be limited with the inclusion of this predictor due to the homogeneity. This is deleted completely.

5. **Feature Creation**

- *Example* : We develop new variables from existing variables to extract hidden relationships of the data. As said earlier, we want to predict the affordabilitty of a home based upon number of baths and half baths. These individual variables may not have as strong of a connection with the response as the total number of baths so we uncover this to improve our future model accuracies.

Conversion of Numeric Variables to Factors(w/levels)

- *Example* : The Porch area in square feet (ScreenPorch) variable was highly skewed to the right (Figure 1), taking on many values of 0 and an extremely small number of positive square feet values. In light of this it is rather important to view this data as simply having a porch or not and convert to it a factor with corresponding levels.

Simplification of Factor Levels

- *Example* : LotShape is simplified into a factor with only 2 levels. The motivation for this change is that there is no need to be too specific with the irregularity of a property shape and it is best classified as irregular or not.

Recoding of NAs

1. **Variable descriptions indicate NA is an absence of a feature**

- *Example* : Basement conditions with an NA indicates having no basement, and they are appropriately filled.

2. **NAs are filled by measures of central tendency / most commonly occuring class**

- *Example* : One observation for the training data's MSZoning (zoning classification of the sale) is an NA and therefore assigned to the most commonly occuring class of RL : Residental Low Density.