Kathy Chowaniec, Li Zhang & Bo Zhang

CS 593: Data Mining II

**Data Mining of Housing Affordability Summary**

For our final project, we decided to analyze the 2013 national dataset by the Housing Affordability Data System (URL: <https://www.huduser.gov/portal/datasets/hads/hads.html>). This dataset is produced from the American Housing Survey conducted by the U.S. Department of Housing and Urban Development (HUD) in order to capture the affordability of housing and to measure the cost burden of housing on households. The dataset contained 64, 535 records with 99 variables of both categorical and numerical types. From this dataset, we selected approximately 20 variables to investigate and include in our regression models. We needed to transform some of these variables using log and square root functions to improve linearity, as well as to create indicator variables for the categorical variables.

The main goal of our project was to use this dataset to predict the current market value of a house/apartment by identifying the main criteria that determines this value. Since this is a continuous value, we chose to use multiple linear regression as the model. We were also interested in predicting affordability by using the cost burden as our criteria to determine whether a house was affordable or not. Since this variable is binary (0 or 1), we chose to use logistic regression as the model in our analysis. From the linear regression of housing market value, we used forward, backward, stepwise and MAXR selection methods to reduce the number of predictors for our target variable. This resulted in a 12-variable model that seemed to have overall significant results. Although we did not see any correlation in our dataset (low VIF values), we also tried using Principal Component Analysis to reduce the number of predictors, however, this method resulted in nine principal components. We finally selected the 6-variable model produced by MAXR as it uses half the original number of variables and there was not much difference in significance, residual plots, or r-square value between the two models.

For our logistic regression model, we used the affordability variable (0 or 1) to try to predict if a house was affordable based on other variables from the dataset. The result of this regression model was that only household income and monthly housing costs remained significant, concluding that these two variables are the most important contributors to a house’s cost burden and affordability. For our final model evaluations, we also split the original dataset into training and test datasets to score each model’s ability to predict the correct values for the target variable. Both of these tests reinforced that our models were significant and seemed to be relatively accurate in predicting the response variable’s value.