IS 590MD

Final Project Proposal

ID: 20

My chosen dataset: Residential Data Set

Linkhttps://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set#

This dataset was donated to UCI Machine Learning Repository on Feb 19, 2018. This data set collected for 360 residential condominiums (3–9 stories) that were built between 1993 and 2008 in Tehran, Iran, a city with a metro population of around 8.2 million (Rafiei and Adeli). Among 29 attributes in this data set, responses are V-9 (actual sales price) and V-10 (actual construction cost) and the rest 26 are predictors. Among predictors, V-1 to V-8 are project physicals and financial variables, V-11 to V-29 are 19 economic variables and indices in 5 time lag numbers.

Data Cleaning:

1. I want to take rolling average of 19 economic variables in 5 time lags because I want to smooth out short-term fluctuations and analyze the variables in a longer term.
2. By subtracting V-10 (actual construction cost) from V-9 (actual sales price), I can create a new variable profit estimated. Profit estimated would be positive if the actual sales price is larger than actual construction cost. In this way, my model can inform potential developers whether he/ she should start a project.

Method:

1. Logistic Regression or QDA. I expect my model to explain 70% of the data.
2. After data cleaning, my data would have continuous predictors and categorical response, where n >>p.
3. I will use k-fold cross validation for model assessment. I will split this dataset into training sets and test sets repeatedly, train the model on training sets and validate it on test sets. I will use MSE to estimate the test error.