Project Proposal

September 23, 2016

Group Members: Bo Zhang, Hanyu Li, Liyun Wang, Zhongyuan Ma

Data from Kaggle.com:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Real estate market is always the one of the most important component of the financial market, and it is very closely tied up to every single individual's life. Everyone inevitably faces the decision to rent or purchase a home at some point of his or her life. However, different from other assets and merchandises, it is far more difficult and takes into account of way more factors when price a real estate asset.

For this project, we consider one real estate class—residential houses, and try to predict the price of each house based on 79 explanatory variables. We obtained the dataset from the "House Prices: Advanced Regression Techniques" competition on Kaggle, and this training dataset is comprised with 1460 entries. A test data set of the same size is also provided. We are going to exploit information from the 79 explanatory variables and build statistical models with this information. These 79 explanatory variables are very basic information about each individual house, and they are easily accessible from the market. They range from general information like the location and the neighborhood of the house to the detailed information such as basement height and electrical system installed in the house. Information like the year constructed, number of bathroom, size of the house, style of the house, garage condition, and many other detailed information are also provided. Basically, these 79 explanatory variables are enough for anyone who have never seen the house to come up with a general picture of what the house is like in his or her head. So with the data that logically justify the price of a house, we believe it is possible to construct a model to predict the price of a house based on these variables.

Since the response variable—housing price is numeric, it is likely that we use regression methods to fit the model. Based on the common logic and understanding of the real estate market, we think that many of the variables will have a linear correlation with the housing price, such as the size of a house. Therefore, it is reasonable to assume that we will extensively use regression methods to construct our models. In addition, since we are given so many variables with relatively not so large data size, we run the potential of over fitting the data with all of the 79 variables. To avoid this, we will also use variable selection tools such as lasso and ridge regression. Many other potential techniques and problems encountered will be discussed in details in this project.

Therefore, being able to price a house based on the available information will bring tremendous value to the real estate industry. As data science and digitalized information are widely available in the era we live in, customers are much better informed about the real estate properties, and are at a better position when negotiate the property price. Likewise, real

estate companies and private companies will also benefit from this model if they are able to have a better estimation on the assets they try to buy or sell. We hope that our model will combine the available information to predict the price of houses, and make a positive impact on real estate industry.