BA 820 Team Project Proposal

Team 8: Huiying Ba, Danting Huang, Ziqin Ma, Jiao Sun, Senbo Zhang

(a) The Problem

The housing market is going through a series of turbulence in recent years, both in the US and the rest of the world. Thus, understanding the factors that contribute to the changes in house prices is more important than ever before. For house sellers and buyers, it is crucial to know the pricing methodology behind in order to evaluate the sale price. For real estate agencies and developers, the analysis we provided will be able to guide them through the process of housing construction and pricing.

In this housing analysis project, we would like to find out the relationships between various housing features and the sale price of the houses in Iowa by using different machine learning models.

(b) The Dataset

The dataset (House Price) comes from Kaggle and has been prepared and split into test and training sets for a case competition (similar to the group challenge in class). The data was collected from Ames, Iowa, with 79 explanatory variables, which can almost describe the relevant aspects of residential homes in that region. The training data consists of 1460 rows and 81 columns while the testing has 1459 rows and 80 columns (excluding the SalePrice column), which in this dataset is the dependent variable we are trying to predict. The goal of this challenge is to predict the house price, using the best fit model generated from the test data. It will be a comprehensive tool to predict the sale price of houses. Below are some examples of the predictors that are interesting to our analysis:

- SalePrice - the property's sale price in dollars.
- MSSubClass: The building class

- MSZoning: The general zoning classification
- LotFrontage: Linear feet of streets connected to the property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of a property
- LandContour: Flatness of the property
- Utilities: Type of utilities available

(c) Proposed Analysis Methodology

Before building any predictive model, we will clean our data by filling in all NA’s with appropriate values. After loading both the train and test datasets, we will split the features into numerical and categorical components.

We will combine supervised and unsupervised machine learning methods to deconstruct our training set, i.e., cluster variables based on underlying features (for example, multi-linearity) and apply regression models (KNN, Random Forest, Boosting, etc.) Focusing on unsupervised techniques, we will try PCA and EFA to group variables and import these components/factors to train supervised models. In practice, we will decide the number of factors based on specific groups and interpretations.

For each trained model, we will compare the Mean Squared Errors and select the one with the least MSE, then import the best fit model to our test set and get the price prediction for each observation. Ultimately, we would be able to provide a reference price for house sellers and buyers on properties with similar features.

Later, we will also experiment with character columns to extract strings that could be factored when clustering.