CS378 Final Project Proposal
Zining Wang, Wenxuan Wang, Wenda Zheng
Kaggle - Housing Prices

Motivation and objective:

People all have their ideal houses. However, when people select and purchase their houses, they need to make tradeoffs between certain criteria and house prices. Every house has its own features such as type of roof, central air-conditioning and etc, and those features all have some impacts on house price respectively, either positive or negative or neither. However, people do not have an efficient tool or standard to estimate the price of their dream houses. Instead, they need to come to see the physical houses and ask the owners the prices, which is very inconvenient. Moreover, the marked price may be higher than the appropriate value. Therefore, the motivation our project is try to find a way to estimate house price based on those features.

Related work and method:

One of the most common methods for this problem is regularized linear regression. A regression model is used to measure the effects of a package of attributes to the house price. In our dataset, 79 house attributes are used as independent variables, and house sales price is the dependent variable. Those 79 house attributes include structural attributes (e.g. the area of first floor), location attributes (e.g. whether the house is adjacent to railroads), and some overall quality ratings. Based on those independent variables, we can build regression models to predict the house price. This will be served as the baseline model for our project.

In addition to our baseline model, some existing methods control the complexity of the regression model by adding a regularized term. The common approach is Ridge or Lasso regularization, and this regularized linear regression usually performed well.

Another common method is XGBoost (Extreme Gradient Boosting). The method is based on the boosting tree model, and predicts the house price based on a sequence of the above attributes. During the training process, it uses gradient descent and boosting method to overcome incorrectly classified subsets over each iteration, until some stopping criterion is met. In this way, the strength of tree learners will be improved. There are existing Python packages for XGBoost, and this method is widely used for data competitions.

What is more, there are also other models, such as neural network or clustering. We can use ensemble learning by combine those classified and weighted their results to make a better prediction.

Proposed Work:

We will first focus on the correlation between individual attributes with the housing prices and find the less interesting attributes that could be excluded from our input data.

Since we are predicting the housing prices from a given data set, we will first try the regularized linear regression. After getting our input attributes, we can use regularized linear regression method to fit our model and predict housing prices. Moreover, since we have structural attributes, location attributes, and overall rating attributes, we can combine many of the original attributes into these categories, add new attributes, and evaluate these attributes based on our algorithm. We can also add some interaction term or dummy variables, to further develop our regression model.

Also, we can try other methods, like XGboost and clustering. After the discretization of the original data set, we can also build XGBoost trees that classifies the range of the housing price based on the given attributes. Another way is to use clustering and cluster the data points based on their attributes. We could use the housing prices within each cluster to predict the housing prices.

After getting all these information, we could analyze the pros and cons of all methods and combine them together based on their correctness.

Evaluation

We will be using the housing prices training data set provided by Kaggle House Prices competition. We will split it into training and testing set. The accuracy of different methods will be calculated by the Root Mean Squared Logarithmic Error (RMSLE), which is the ranking standard by Kaggle. Our goal is to minimize this error. We will also submit to Kaggle, and test on unseen testing data.

Plan of action:

There are four weeks remaining for our project. We plan meet at least twice per week. On every meeting, we discuss our progress, and then assign further tasks.

Week1: Preprocess raw data and analyze the correlation between individual attributes with housing prices.

Wenda will work on the data preprocessing and will clean the data. Zining and Wenxuan will find the more influential attributes among all 79 house attributes that will help reduce the complexity and improve the accuracy.

Week2: Try regularized linear regression and analyze the results.

We will work altogether on the same thing but may have different focus. Zining and Wenda will use packages to perform the regularized linear regression. Wenxuan compare the result with the expected ones.

Week3: Try XGBoost and other models and analyze the results.

Zining and Wenxuan will work on the code. Wenda is assigned to perform analysis.

Week4: Compare and evaluate the results of two methods

We will meet and compare the results obtained from different methods and calculate their reliability and accuracy. A final method to combining our result will be proposed and applied to our final prediction.