# Independent project report

Jaeha Huh

## 1. Goal

The goal of the project is to divide the dataset into a training set and a test set, and predict housing prices using a supervised machine learning models with given features (column).

## 2. Data structure

The dataset has a total of 81 columns, including street, house size, building type, house condition, number of toilets, and other factors used to evaluate the house. Columns were divided into nominal_vars, ranking_vars, and continue_vars, respectively. In the case of ranking_vars, the values are strings, and each value is a column representing a ranking or step. In the case of continue_vars, the values are columns that are numeric columns. Columns that do not fall under these two were included in nominal_vars.

Ex)

nominal_vars = 'MSZoning', 'LandContour', 'Utilities', 'LotConfig', 'Neighborhood', 'Condition1', 'Condition2', 'BldgType', 'HouseStyle', 'RoofStyle', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'MasVnrType', 'Foundation', 'Heating', 'Electrical', 'GarageType', 'MiscFeature', 'SaleType', 'SaleCondition'

ranking_vars = 'OverallCond', 'ExterQual', 'ExterCond', 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2', 'HeatingQC', 'KitchenQual', 'FireplaceQu', 'GarageQual', 'GarageCond', 'PoolQC', 'Fence', 'Street', 'Alley', 'LandSlope',

'Functional', 'GarageFinish', 'MoSold', 'YrSold', 'PavedDrive', 'CentralAir', 'LotShape', 'MSSubClass',

continue_vars = 'LotFrontage', 'LotArea', 'MasVnrArea', 'BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'BsmtHalfBath', 'FullBath', 'HalfBath', 'TotRmsAbvGrd', 'BedroomAbvGr', 'KitchenAbvGr', 'Fireplaces', 'GarageCars', 'GarageArea', 'WoodDeckSF', 'OpenPorchSF', 'EnclosedPorch', '3SsnPorch', 'ScreenPorch', 'PoolArea', 'MiscVal', 'GarageYrBlt', 'YearBuilt', 'YearRemodAdd', 'OverallQual'

3. Missing Values

There is a significant amount of missing values in this dataset. The method of deleting missing values is also one way, but in some columns, more than 90% of the data was found to be missing values, so the method of deleting them was not used. The data types in each column are also different, so I filled in the missing values using a replacement method suitable for each column. For columns with string data, None is inserted for missing values, and 0 is inserted for columns with numeric data. In this case, mean value imputation was used for columns that significantly affect model training when replaced with 0 or None. Finally, the string data is replaced with a numeric data using the get_dummies() method. In this case, the method used by get_dummies() is One-Hot-Encording.

Ex) LotShape: General shape of property (This column does not have missing values and it is ranking_vars)

Get_dummies() : One-Hot-Encording

| ID | Reg (Regular) | IR1 (Slightly irregular) | IR2 (Moderately Irregular) | IR3 (Irregular) |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 0 |
| 3 | 0 | 1 | 0 | 0 |
| 4 | 0 | 1 | 0 | 0 |
| 5 | 0 | 1 | 0 | 0 |
| 6 | 0 | 1 | 0 | 0 |
| 7 | 1 | 0 | 0 | 0 |

As a result, the number of columns of data increases.

4. Machine learning models

Models applied to supervised learning tested a total of three models: MLP Regressor, Ridge Regressor, and XGB Register. After testing normalized and unnormalized data, proceed with a model that can achieve better results.

1) MLP Regression is short for multi-layer perceptron, which can be described as an upgraded version of logistic regression and artificial neural networks. MLP has the advantage of being able to better divide the division boundary in the form of adding a hidden layer between the input layer and output layer, which are logistic returns.

2) Ridge Regression is a model that uses a normalization method to increase the predictive power of a linear model. Using the basic linear model is very suitable for frequently

occurring overfitting, i.e., data, resulting in extremely fluctuating graphs, and the coefficient value of linear regression representing it is large. To prevent this situation, ridge regression is a random small adjustment of the coefficient by adding an equation. Therefore, it is a technique that can expect good results by minimizing errors and penalizing functions. In the case of this dataset, after confirming that there are as many as 80 columns and that the coefficients of each value are large, I chose ridge regression.

3) In the case of XGB Regression, it is an ensemble method based on a decision tree, especially a machine learning technique based on a boosting method. It is also introduced as a good technique to prevent overfitting and has the advantage of supporting cross-validation cross-validation. Thus, I chose XGB Rregression,