Project Proposal

Group 15

Context:

Real estate pricing has been a challenging problem for decades: size, location, year-built, or even a basement can become a parameter that influences the real estate's price. Hence, in most circumstances, the buyer and the seller can only price the house subjectively. With the development of machine learning and deep learning technology, our team is trying to find out if we can use machine learning regressor to give a precise and objective price for real estate. Hence, we choose to use multiple machine-learning regressors to predict the house price and find out which machine-learning regressor has the highest accuracy in predicting house prices.

Data Set:

Link: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

This dataset is from an ongoing Kaggle competition. The author of the dataset, Dean De Cock, collected the 2930 sale records of individual residential property in Ames, Iowa. (Collecting date range from 2006 to 2010) Data Set has 80 variables (including the target variable: SalePrice): 20 continuous variables, 14 discrete variables, 23 ordinal variables, and 23 nominal variables. With the python pandas package, our group found out that there is no missing value within the dataset. For more details about the data set, please viewed https://jse.amstat.org/v19n3/decock.pdf

Data set pre-processing:

For continuous variables, we will use a co-relation matrix for variable selection (for high correlation variables, we will only keep 1 variable); For discrete variables, we will just keep them; For ordinal variables, most of them are binary. Hence, we will transfer them to 0/1 discrete columns; For nominal variables, we will use one-hot encoding to encode them.

Train Test Split:

We will split the raw dataset into 60% training, 20% validation, and 20% testing.

Machine learning algorithm we will use in our project:

Benchmark: Linear Regression

The algorithm we will tune:

Ridge Regression, Lasso Regression, Random Forest Regression, Support Vector Machine Regression(with GridSearchCV), Multi-layer perceptron Regression, and XGBoost Regression.

Others:

Seaborn and matplotlib for Data Visualization.