CS 559 Final Exam Part 2 Description
Due 5/16th Saturday 12 PM

Objective:
You are going to predict the house price. Thanks to a large number of features in the dataset, you are going reduce the train data set and make a simplest model you can make to predict the house price.

Exam Rules and Requirements:
1. Using the train data set, determine the categorical and numerical variable columns. [6 pts]
    a) Determine the shape of data set. [3 pts]
    b) How many categorical features do you have? How about numerical features? Do you think you can eliminate all string data? [3 pts]
2. Using the train data set, determine features having missing values. [3 pts]
    a) Select top 5 features having the most missing values for both categorical and numerical features.
    b) Provide the feature names and their missing value rates (# of missing row/3 of total row). [3 pts]
    c) Make sure to read the data description carefully because you will be using some of them.
3. Using the train data set, do feature engineering and EDA. [24]
    a) Among categorical features, did you notice any features can be eliminated? What are they and explain. [3 pts]
    b) Find the correlation between variables. Did you notice any tight correlations and possibilities to eliminate them? You can visually examine them.
        a. Do this for Sale Price > 0.5 and ≤ 0.5 [3 pts]
    c) Among numerical features from 2-b, select 3 of them and fill them. [3 pts]
    d) Reduce the train data set to 40 features or less. In here, **you must include the three features you filled in missing values**. [15 pts]
        a. To do the feature selections, you can do stepwise regression, forward selections, and backward elimination or k-fold or clustering.
        b. You can do other way to make decision.
        c. Display first 10 rows of data frame after step d.
        d. Explain your workflow and decision.
4. Up to this point, you will be good to make a final model. [20 pts]
    a) Using a finalized train data set, you are going to use make three different models.
        a. Linear Regression or Multi-linear regression.
        b. Random Forest or AdaBoost/Gradient Boost/XGBoost (if you used Random Forest in step 3, use Boosting model)
        c. Combined models
        d. Calculate the RMSE and provide results for all models you trained.
        e. Repeat step 3 and reduce the train set to its 80%.
        f. Repeat step a to d.
        g. Write a short explanation on which model will be used to make the final prediction.
5. Select the best model among three models you have in 4 and make a final prediction using the test data set. [8 pts]
    a) Calculate the predicted value and RMSE of final model. [3 pts]
    b) For each observation, give value 1 if the actual target value agrees to predicted value within RMSE otherwise give 0. The accuracy will determine by the RMSE value and the counts of 1.
    c) Top 2 highest accuracy scores [5 pts], next two [4 pts], rest [3 pts]

Data Description:
There are three files: data description, train, and test.