CISC-351 Advanced Data Analysis
Assignment 2

February 2, 2019

Given the house price data, our goal of assignment 2 is to build a regression model for house price prediction. This task should be performed in two steps:

1 Problem 1

Preliminary data analysis:

- Analyze the distribution of the house price in the given (training) data set. Explain what do you find.
- Analyze correlation between each feature and house price. Determine the features you think may be useful for the house prediction task. If you think you should perform any feature transformation, describe what kind of transformation you want to apply and explain the reason.

2 Problem 2

Using the given data set, based on your selected features in section 1, create three models (random forest, gbm, xgboost) to predict house price and select appropriate metrics to compare the performance of the three models on the testing data. Explain how you determine the input parameters of your models and describe what are the important features given by each model.

3 Assignment Requirements

Answer the above questions, provide figures to explain your answers. R markdown file(pdf) is highly suggested. However if you are still not comfortable with R, you may try Knime. You might want to check this reference: `https://www.kaggle.com/kyen89/house-prices-prediction-with-ensemble`
Deadline: Feb-16 12:00am, 2019
Submit: OnQ in pdf format.