Capstone Project 1 Proposal by Hye Joo Han

1. Overview

In this project, I will find a model that predicts house sale prices the best using various machine learning techniques. Beelow (imaginary firm), a local real estate company serving King County in Washington, asked to build a house price prediction model. The company wants to utilize the model to find good estimated house prices to serve house sellers and buyers.

2. Dataset

I will use a dataset, House Sales in King County, USA, on Kaggle.com (Link: https://www.kaggle.com/harlfoxem/housesalesprediction). This data contains house sale prices for King County included in the Seattle–Tacoma–Bellevue metropolitan area in Washington. FYI, house prices in the Seattle metropolitan area are rising the fastest in U.S. The dataset has only one year of records for houses sold between May 2014 and May 2015. Thus, I will assume that this project is being done in 2015. The dataset consists of 19 house features, house id and sale price (target) for 21,613 observations. The house features include the number of bedrooms and bathrooms, square footages of home and lot, year built, zip code, latitude and longitude. Some other factors such as public schools and safety can have a big impact on the house prices. Therefore, I will check if I can gather more features related to the environment of house using the latitude, longitude or zip code features. For example, some websites like www.niche.com or https://www.kingcounty.gov/services/data.aspx could be useful.

3. Method

1) *Data wrangling:* I will combine datasets from different sources if I can find some useful extra datasets. Then, I will clean and transform the datasets if necessary.
2) *Exploratory Data Analysis (EDA):* I will do EDA to understand and summarize the dataset and get insights. I will first check missing values and outliers. Some univariate, bivariate and multivariate analyses are to be done with visualizations. I will also use Inferential statistics to obtain more rigorous findings if useful.
3) *Machine learning:* I will try several machine learning algorithms (e.g., Random Forest, XGBoost, LightGBM, or Neural Networks) or their ensembles and find the best model. The performance measure that evaluates models will be carefully selected.

4. Deliverables

This project will be done using Jupyter notebook (to be shared) and the main findings will be shared through presentation slides and technical report.