Capstone Project 1 Proposal

Prediction of House Prices in King County

1. The Problem

What is the problem you want to solve?

House market is a big yet complex market and is often affected by the overall economy, policies and neighborhood, making house prices difficult to predict. Besides, information asymmetry in the house market makes the prediction of future house prices more challenging.

Thus, the application of advanced data analysis and machine learning plays an important role in house prices prediction. The goal of this project is to apply machine-learning models to predict future house prices with accuracy and classify houses in order to provide a healthier house market by reducing information asymmetry with the application of data science.

2. The Client

Who is your client and why they care about this problem?

Residential real estate companies and real estate investment firms that are looking for higher Return on Investment (ROI) will be interested in this project. With the application of the prediction model, companies have predicted house prices and know what kind of house they are going to invest, and thus reduce the risk of investing in the wrong house market.

3. The Data

What data are you going to use for this? How will you acquire the data?

The data contains house prices in King County, WA, USA from 2014 to 2015. The dataset has more than 200,000 rows, 19 attributes and has numeric and date time data type. Dataset:
https://www.kaggle.com/harlfoxem/housesalesprediction/data

4. Solving Problem Approach

Outline your approach to solving this problem

Initially, we will focus on data cleaning and data wrangling, including transforming data type accordingly, and dealing with missing data and combining different datasets for analysis.

The second part will contain exploratory data analysis to obtain a better understanding of our data. We will use data visualization to see how the house price changes in King County and plot out where are those houses located.

The final part will apply machine learning with statistical analysis. We will apply supervised machine-learning model to identify important features using regression model and to predict future house prices. In the last part, we will test the prediction model to make sure the model predict house prices with accuracy.

5. The Deliverables

What are your deliverables?

The deliverables will be

- Report: The report will include Business Understanding, Data Preparation, Modeling & Evaluation and Deployment. This will help clients better understand the objective of this project and the approaches we use.
- Slide deck: Using slides to present the analysis that people can easily understand and relate to.
- Source code: GitHub repository and a documented code help clients with technical background find the information they need conveniently.