Machine Learning Engineer Nanodegree

Capstone Proposal

KT Wu

2018/4/15

Proposal

Domain Background

In the case of rising wages, there are still many Taiwan young people who simply can't afford to buy a house.(Taiwan house price problem is hot!) Otherwise, they have to bear 20 or 30 years of mortgages. After the government publishes maps of various disaster potentials, it is feared that they will have to lose money. However, people say that although the process of buying a house may not necessarily make money, the process may still be very painful. It is necessary to "require the daily necessities", and it is necessary to buy one when people are young as soon as possible.
**Lalaine C. Delmendo. Taiwan's housing market is gaining momentum from http://www.globalpropertyguide.com/Asia/Taiwan/Price-History [Aug 15, 2017]

Problem Statement

Many people seeing at the house and buy a house. They all have a question in their minds: "What is the reasonable price of a house?" When we really like a house, what price we want to pay will not be too high. In Taiwan, real estate information is not so easily to get, people always "heard" from the housing agency but they just want you pay for more agency fees, so how to pay for reasonable price is a problem which could solved by machine learning.

Datasets and Inputs

In order to practice the housing price problem, I download the data from Kaggle. The

training data contains 80 features and data set split as two parts: 1460 training data(include some blank entry, so data preprocess is needed.) and 1459 test data.
- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- Neighborhood: Physical locations within Ames city limits.
- LotArea: Lot size in square feet
- YearBuilt: Original construction date

**Kaggle https://www.kaggle.com/c/house-prices-advanced-regression-techniques/download/train.csv

Solution Statement

Consider the nearest house price would be a reasonable solution. KNN the strait way which predict by grouping the nearest data point. House price just easily suit the algorithm. In the same street or block or building, there house price would be closed in unit cost, or maybe combine the floor number or some factors. so decision tree may also be a good choice of prediction for the multi level feature combination to the house price problem.

Benchmark Model

In my first thought, "location, location, location" is the golden rule in many ways of real estate, so the neighborhood house price of the same location is most important feature of house price prediction model. If I have house price of up/down floor then I could predict the price would be mean of the up/down floor.

Evaluation Metrics

- Final sales price is the final index of this problem, when a new or old house is going to be sold, the sale price is the first and only thing most people care, so the price is a reasonable way to evaluate the model.
- Unit price is also the objective index, so many people could predict total house price by easily times floor size and unit price price.
- parking lot are usually an issue of house price in the city, parking lot unit price sometimes higher than floor unit price. Final price = (Floor space) x (unit price) + (parking lot price) For the house price prediction model, the smaller RMSE would be a better model.

Project Design

The first step is downloading the dataset from anywhere like Kaggle or other pipeline.

Review the dataset and simply take few of this dataset to try the models of previously mentioned (Average the neighborhood, KNN, Decision tree). Preprocessing the dataset is always needed, to remove the uncompleted data or normalize dataset. Trining model with training data and do the cross validation to prevent overfitting. Test the model with new data (test data) to verify the model is workable or not. Review the difference between prediction result and real result to optimize the model.

Before submitting your proposal, ask yourself. . .

- Does the proposal you have written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Solution Statement** and **Project Design**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your proposal?
- Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?