Machine Learning Engineer Nanodegree

Capstone Project

K.T. Wu

April 22st, 2018

I. Definition

Project Overview

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. It's an issue to estimate house price for the house information is not so clear. I think machine learning will be helpful to people who want to buy a house with but don't know how to estimate the price. combine all the information (features) to estimate the reasonable house price. 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.

Problem Statement

People who want to buy a house always encounter a problem: How to offer the reasonable price? base on many many information like location, house size, room numbers, parking lot, built years, public utilities, convenience store…etc. and the house price is the last result to present.
- Download dataset from Kaggle, it is a summarized dataset with 80 features like mentioned before and the final result is SalePrize.
- Preview the dataset with some statistic analysis.(mean, median, data counts…)
- Filter the outlier.
- Normalize huge variation items.

- split dataset as train and test data.
- fitting data with kNN, decision tree,
- RMSE is the final KPI to evaluate models.

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.

II. Analysis

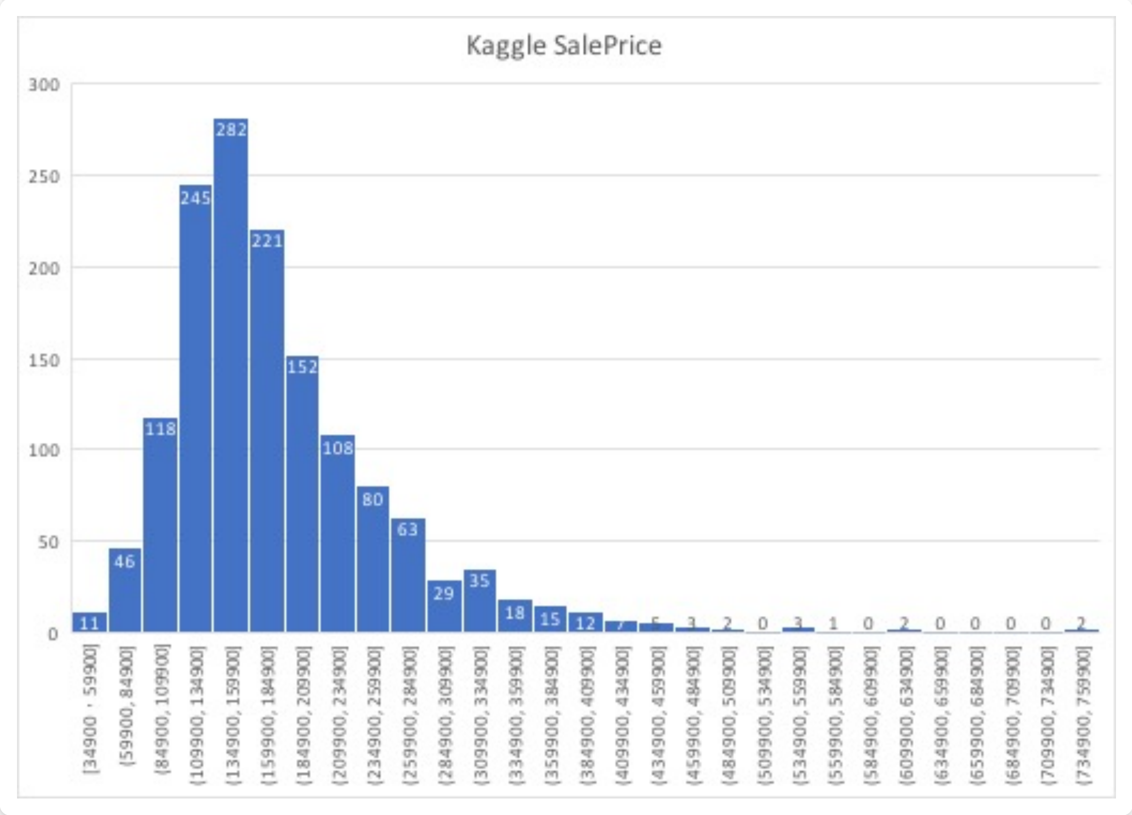
(approx. 2-4 pages)

Data Exploration

The house price dataset comes from Kaggle, with 1460 real estate sale price as dataset. There are total 80 kind of house information like size or room numbers and include neighborhood in string type content.
- SalePrice (float)- the property's sale price in dollars. This is the target variable that you're trying to predict. (from 34900~755000, 20X data range, median is 163000 and mean is 180921)
- Neighborhood(string): Physical locations within Ames city limits.
- LotArea(float): Lot size in square feet(1300~215245, 150X data range)
- YearBuilt(integer): Original construction date

Exploratory Visualization

From the bar chart of SalePrice reveal the major sale price is in 134900~159900 , so if the best guessing is in this range. And the average is 180921 in the next bar (159900~184900) that means the average could be affect by some extreme value, we can see the maximum price comes to 755000 which is 4times than average also has the
.



Algorithms and Techniques

I use decision tree as classifier which is highly explainable model in machine learning.
from sklearn, decision tree regressor has 12 parameters to manipulate:
- criterion='mse': The function to measure the quality of a split. Supported criteria are "mse" for the mean squared error
- splitter='best'
- max_depth=None,: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
- min_samples_split=2, :The minimum number of samples required to be at a leaf node(i.e. 2samples)
- min_samples_leaf=1,
- min_weight_fraction_leaf=0.0,
- max_features=None,
- random_state=None,
- max_leaf_nodes=None,
- min_impurity_decrease=0.0,
- min_impurity_split=None,
- presort=False

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

Benchmark

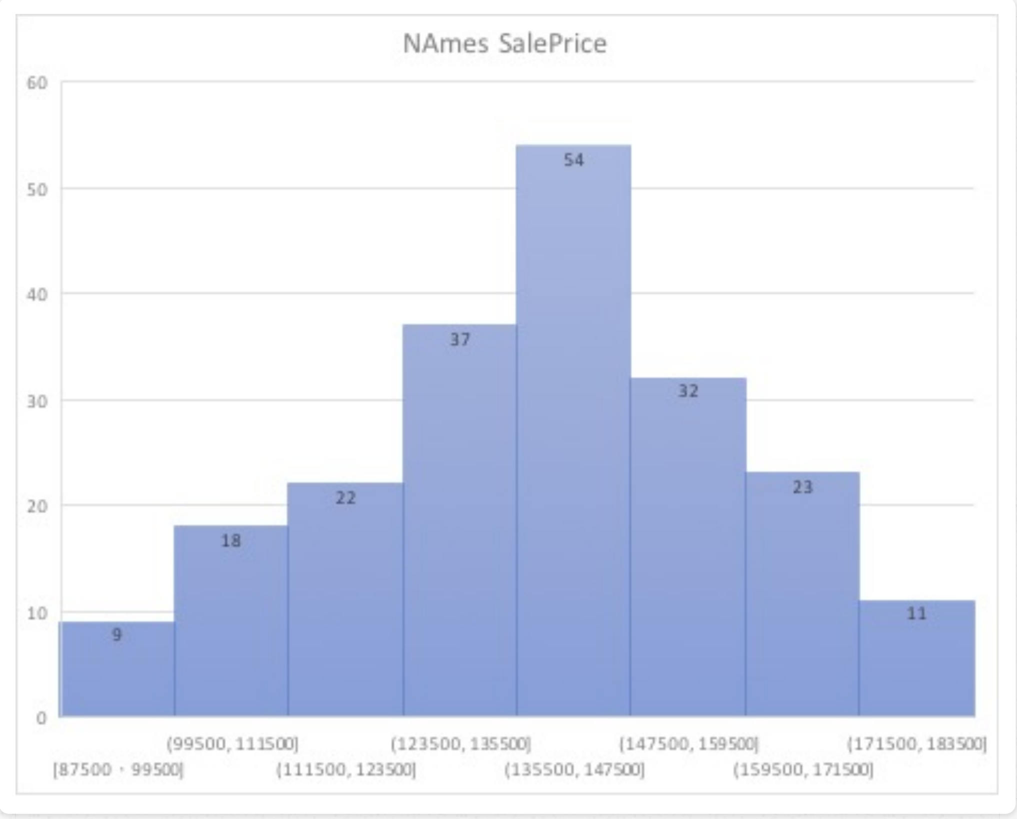
In my first thought, "location, location, location" is the golden rule in many ways of real estate, so the neighborhood house price of the same location is most important feature of house price prediction model. If I have house price of up/down floor then I could predict the price would be mean of the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

III. Methodology

(approx. 3-5 pages)

Data Preprocessing

- Remove or refill NA data: some features(ex, LotFrontage) has "NA" in a number column, like 'LotFrontage'
- Grouping by 'Neighborhood': as I mentioned, the house price is highly correlate with local area. I filter the 'NAmes' with 225 data as a group to find the model and it seems more converge.



Implementation

- I choose the 'Neighborhood: NAmes as analysis target area.

- split all data as train and test dataset with 80/20 ratio.
- Import DecisionTreeRegressor from sklearn.
- train model with data_model.fit()
- test data with data_model.predict()
- calculate the mean absolute error of prediction and test_y: 20925

Refinement

I tried first is through all the dataset into decision tree model and the mean absolute error result was up to 29332. So I narrow down to neighborhood 'NAmes' as new dataset and re-train the model. I got the mean absolute error: 20925 is reduce 30% error than before and which is in the decision tree default parameter setting, so I extend the max_leaf_node from 'None' to [2,3,5, 10, 100, 1000] to find the minimum error and the result:
Max leaf nodes: 2 Mean Absolute Error: 15465
Max leaf nodes: 3 Mean Absolute Error: 15048
Max leaf nodes: 5 Mean Absolute Error: 16104
Max leaf nodes: 10 Mean Absolute Error: 17262
Max leaf nodes: 100 Mean Absolute Error: 18674
Max leaf nodes: 1000 Mean Absolute Error: 18818
So I know the better setting is Max_leaf_node = 3, it is huge reduce the model almost 40% error!

- _Has an initial solution been found and clearly reported?_
- _Is the process of improvement clearly documented, such as what techniques were used?_
- _Are intermediate and final solutions clearly reported as the process is improved?_

IV. Results

(approx. 2-3 pages)

Model Evaluation and Validation

The decision tree model perform the thinking of my consideration and the evaluation of max_leaf_node is surprising good to refine the model. To find the best setting I use a list [2,3,5,10,100,1000] and the final setting max_leaf_node = 3 would let the model being explainable.

Justification

As we can see from mean absolute error is 15048 in the final condition (localize to NAmes and the max_leaf_node = 3) which improved almost 100% of model prediction ability from the previous guessing model in all dataset (mean absolute error = 29332) although it is suitable for localize house sale price however in practice, we only focus the neighborhood we interested.

V. Conclusion

(approx. 1-2 pages)

Free-Form Visualization

In the fig.1, all the dataset is plotted in one chart, and the data is shifted to the left side for some extremely point (extremely high), which is one of the reason that affect the model prediction ability.

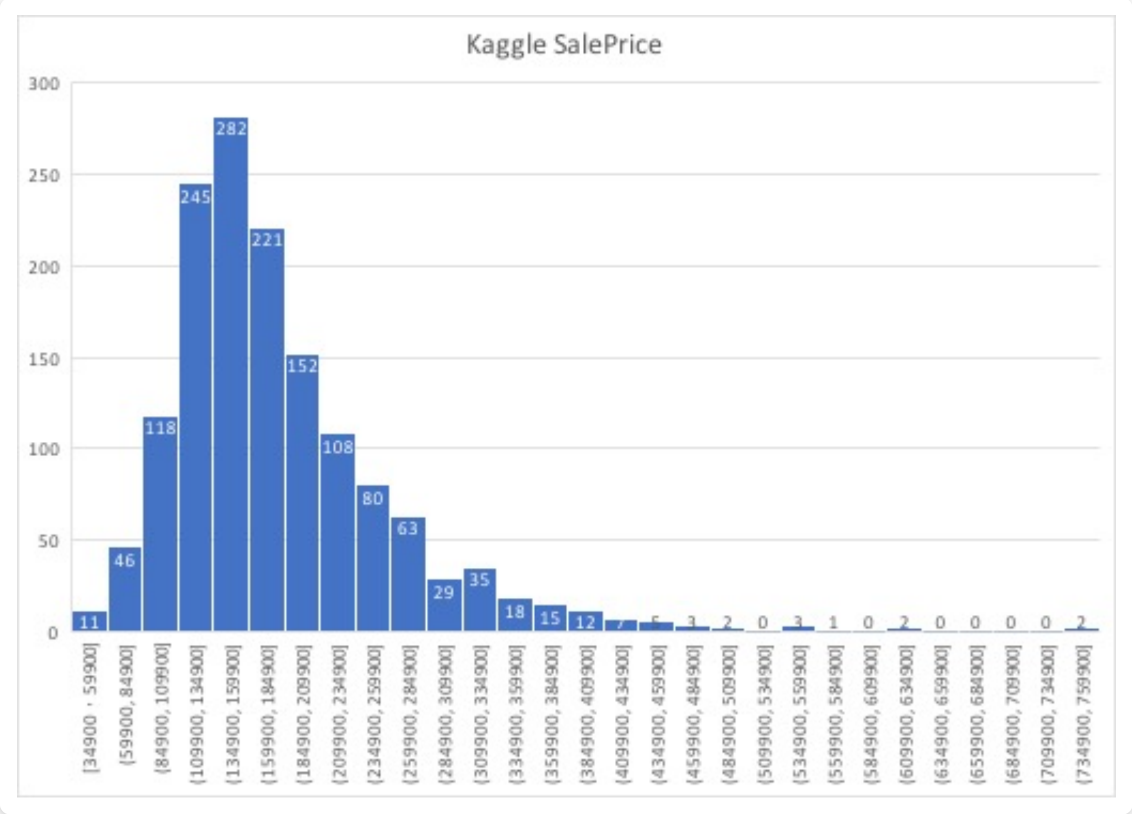


Fig.1 All SalePrice

So I narrow down the dataset to neighborhood 'NAmes' and remove the extremely high point so the chart is more converge and centralized.

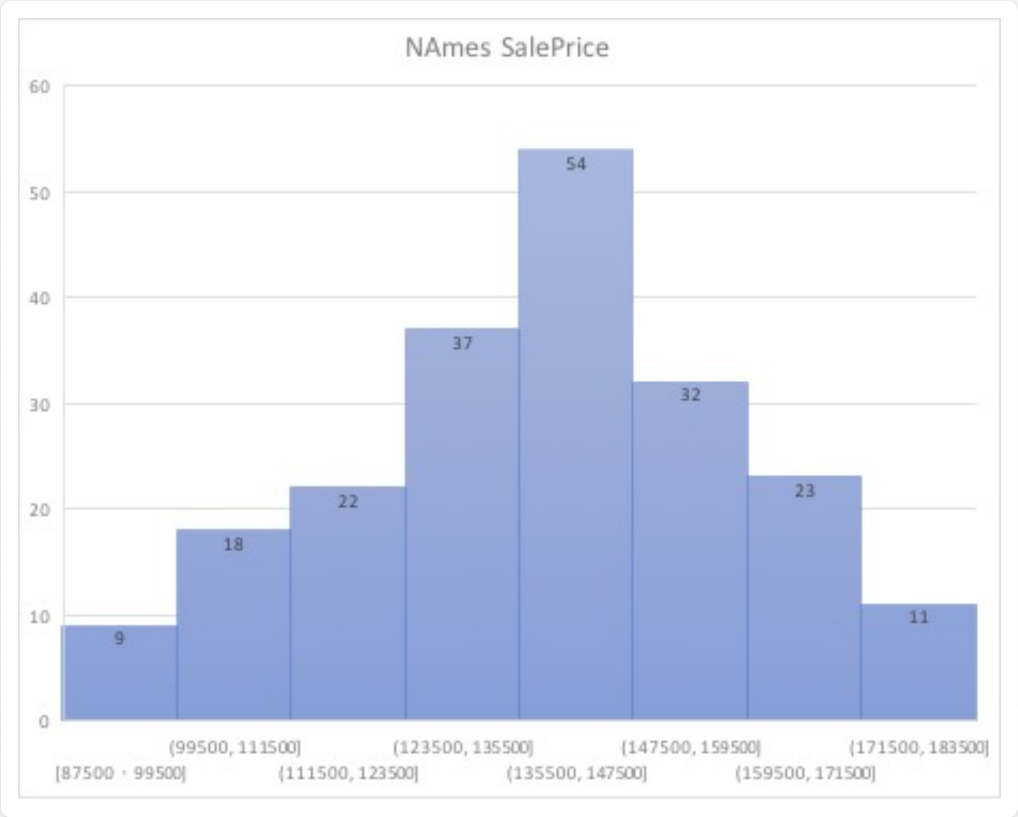


Fig.2 'NAmes' SalePrice

Reflection

From beginning of the topic, the house sale price is an truly issue in real world. In many ways people often misguide by the real estate agent so I think machine learning model could help to estimate house price objectively.
I found the dataset from Kaggle and preprocess some null data of features (ex. lotfrontage) then I split entire dataset to training and testing dataset.
First I through the entire dataset to the decision tree model to train and the mean absolute error is 29332 then I narrow down the neighborhood 'NAmes' and test the parameter max_leaf_node with a list from 2 to 1000 then test the model to find to minimum value of mean absolute error 15048. I believe the local area house sale price is more meaningful than global house sale price.

Improvement

I think there may have a technique to improve: a Global house sale price predictor. As I narrow down the only one neighborhood 'NAmes' means house sale price of other neighborhood are dropped. One-hot encoding maybe an approach but when the neighborhood goes larger the dataset must be re-process again so it is not so logical solution. Alternatively, other model may have better performance to fit the dataset.

Before submitting, ask yourself. . .

- Does the project report you've written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Analysis** and **Methodology**) 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 analysis, methods, and results?
- Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?
- Is the code that implements your solution easily readable and properly commented?
- Does the code execute without error and produce results similar to those reported?