Housing Prices Prediction Project

Rational Statement:

The prediction of property prices for real estate markets is becoming increasingly important and beneficial. Property prices are a good indicator of both the overall market condition and the economic health of a country. Housing price ranges are of great interest for both buyers and sellers.

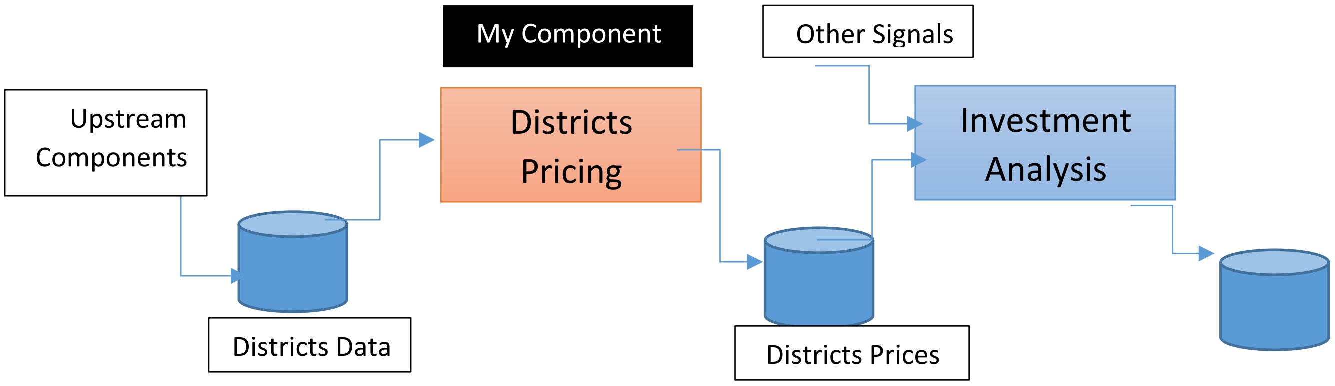
Motivation:

In this project, As a first experience, I want to cover many data science concepts such as data cleaning, feature engineering, dimensionality reduction, outlier removal…etc. as much as possible by approaching every different steps of the machine learning process and trying to understand them deeply. I chose to take Real Estate Prediction as approach. The goal was to predict the price of a given house according to the market prices taking into account different “features” that will be developed in the following sections.

Problem Statement:

The real estate company estimate the houses prices manually by experts who collect data about the district to study it and apply complicated rules to get the median houses prices. This procedure consumes much time, much money and much effort. In addition, there is fairly error percentage.

The company needs a data scientist to build a machine-learning model, which can predict houses prices in districts with low error percentage. This model will go with other signals to feed another ML model that predicts whether is good to invest in the area or not.



Frame the problem:

After discussing with the business manager, we got this information.

- The type of machine learning that will be used is Supervised Machine Learning. Because the training data are labeled and contains the desired class which is the median houses price.
- Because the desired feature that we will predict is value, the task of the machine learning will be Regression.
- The data is not streaming. Therefore, we will use batch learning.
- As the model will predict house price based on many features, the regression type will multivariate regression problem.

Select a performance measure:

As the problem is regression task. Therefore, I will use the most popular function for that which is RMSE (Root Mean Square Error). We use RMSE to measure standard deviation.

$$RMSE = \sqrt{\frac{1}{m}\sum_{1}^{m}(h(x) - y)2}$$

M: number of instances.

X: vector represents all features except the desired feature.

y: the desired solution label. (house price).

h: the hypothesis function.

Data Requirements:

Data Source:

I will use the California Housing Prices dataset(housing.csv) from the following kaggle site:

https://www.kaggle.com/camnugent/california-housing-prices.

The dataset contains 2064020640 observations and 10 attributes (9 predictors and 1 response). Below is a list of the variables with descriptions taken from the original Kaggle site given above.

- longitude: A measure of how far west a house is; a higher value is farther west
- latitude: A measure of how far north a house is; a higher value is farther north
- housing_median_age: Median age of a house within a block; a lower number is a newer building
- total_rooms: Total number of rooms within a block
- total_bedrooms: Total number of bedrooms within a block
- population: Total number of people residing within a block
- households: Total number of households, a group of people residing within a home unit, for a block
- median_income: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
- ocean_proximity: Location of the house w.r.t ocean/sea
- median_house_value: Median house value for households within a block (measured in US Dollars)

Data Assumptions:

To make sure about the understanding of the machine learning task, I need to ask the work team who work on the next machine learning component to understand their needs. As they need the actual values of the houses prices, I will select the regression task in my work.

Limitations:

The dataset is about houses data in California in 1990. I am fully aware that housing prices have increased dramatically since 1990, when the data was collected. This model should not be used to predict today's housing prices in California. This is purely an academic endeavor to explore statistical prediction.

Test Process:

I will split the data into three classifications i.e. training dataset, validation dataset and test dataset. The model will be trained using training dataset and then evaluate using validation data. Based on accuracy, I will retrain the model until I get an acceptable accuracy.

there are also a number of variables that are worth looking into. Therefore, I choose to study the house prices predicting problem on Kaggle, which enables me to dig into the variables in depth and to provide a model that could more accurately estimate home prices. In this way, people could make better decisions when it comes to home investment.

House prices will be predicted with various regression techniques including Lasso, Ridge, SVM regression, and Random Forest regression. I will also perform PCA to improve the prediction accuracy. The goal of this project is to create a regression model and a classification model that are able to accurately estimate the price of the house given the features.