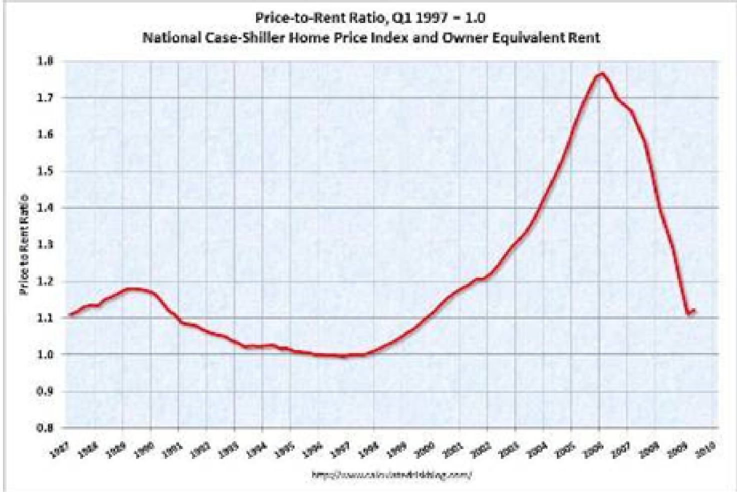
## EE4483 Artificial Intelligence and Data Mining

# Mini Project (Option 1)

**Due: Friday (11:59pm), 28 July 2023**



US Median Price of Houses Sold
[Source: https://en.wikipedia.org/wiki/File:U.S._Housing_Price_Measures_-_Index_and_Dollar_Price_Value.png]

You are required to build a learning-based algorithm to predict housing prices, based on historical training data containing several housing features and sale prices. The whole process involves the following steps:

1. **Reading in data**: You will need to load the data from the provided csv files. There are two csv files, namely training set ("train.csv") and testing set ("test.csv"), each contains examples in the dataset splits. The attributes of the data are as follows:
   a. Id
   b. BldgType - Type of the Building
      There are 5 types: 1Fam, 2FmCon, Duplx, TwnhsE, TwnhsI
   c. OverallQual - Overall Quality
   d. GrLivArea - Ground Living Area
   e. GarageArea - Garage Area
   f. SalePrice - Sale Price (to predict in testing)

   The "ground truth" for the "SalePrice" of each sample is provided in the training set.

2. **Data processing**: You will need to convert the raw data into appropriate feature format. For example, you can convert the BldgType into a one-shot vector of length five, as there are 5 types in total. If one sample has BldgType = '1Fam', it can be converted to [1, 0, 0, 0, 0].

3. **Model Selection**: You are going to conduct a regression model to predict the house sales price for the testing data. For example, $F = f(x)$, where x is your input feature, F is the predicted price, and f(.) is the algorithm that you designed. You can use **neural networks** for this task. Backpropagation (BP) can be applied to compute the gradient of the model parameters and update the model.

4. **Training and Tuning**: You will need to try different model parameters to obtain good regression results, e.g., feature dimension, weights, initialization, and learning rate. You can use the training set for model training and tuning. In the training stage, your algorithm should take as input only the attributes available in the training set, excluding attribute "SalePrice", which is the output of the regression algorithm.

5. **Prediction:** By setting up the correct learning algorithm, you can predict the sale price of the samples accurately for the testing data. You need to submit the predicted "SalePrice" for the testing data in the file ("submission.csv").

Submit your report to **answer the following questions**:

(a) Understand the training and testing datasets that we provided: Make a table to describe the two datasets, including their feature dimension, number of samples, mean and variance of the values in each attribute. (5**% marks**)

(b) Select at least one appropriate model (e.g., linear regressor, neural network, support vector machine, etc.) to build your regressor. Clearly describe the model you use, including the input and output dimensions, structure of the model, loss function(s), training strategy, etc. Include your code if you are solving the problem by programming. (**15% marks**)

(c) Discuss how you consider and determine the parameters (e.g., learning rate, etc.) / settings of your model as well as your reasons of doing so. (**15% marks**)

(d) Apply the regressor(s) built to the test set ("test.csv"). Submit the "submission.csv" with the results you obtained. (**20% marks**)

(e) How many of these 259 testing samples are predicted to be more expensive than 250000? Among these expensive samples, how many of them have (i) GarageArea > 700, (ii) GrLivArea > 2000, and (iii) OverallQual > 8? (**10% marks**)

(f) What is the most dominant factor for high housing price in your model? Why? Please justify your conclusion based on the prediction you made. (**10% marks**)

(g) Build one more regressor (either a new model, or your existing one with different loss functions), discuss and compare the results obtained from the different regressors. (**10% marks**)

(h) Extend your regression algorithm, or build another network (e.g., CNN) for image classification for **Cifar-10** dataset (https://www.cs.toronto.edu/~kriz/cifar.html). Describe details about the dataset, classification problem that you are solving, and how your regression algorithm can be modified to tackle this classification problem. Report your results for the testing set of **Cifar-10**. (**15% marks**)

[Hint: an image can be considered as a matrix. If you wish to extend your algorithm for images, you can vectorize an image matrix into a vector, and apply PCA to reduce the vector dimension to fit your regression network input.]

Notes:

- You can choose any programming language / platform that you like to complete the task.
- If you couldn't obtain any meaningful results or answers to the questions above, you may describe what you have done and attach the relevant working, codes, or screenshots, if available.
- You should clearly cite all the references and sources of information used in your report.
- You are expected to uphold NTU Honour Code.
- Submit your report and the file "submission.csv" with your results to the assigned TA via NTULearn by the **deadline**: Friday (11:59pm), 28 July 2023.