# CPE695 Final Project Mid-Stage Report

Yupeng Cao
*ECE Department*
*Major in Applied Artificial Intelligence*
10454637
ycao33@stevens.edu

Haixu Song
*ECE Department*
*Major in Applied Artificial Intelligence*
10446032
hsong13@stevens.edu

Chong Guo
*MA Department*
*Major in Data Science*
10451296
cguo10@stevens.edu

***Abstract*—This document is mid-stage report for the final project in CPE695 Applied Machine Learning course. The project is trying to implement a housing price prediction model by using varied machine learning algorithms. The housing price prediction problem is defined in first part and we introduced the American Housing Survey (AHS) data set. In mid-stage, we finish data pre-processing step and utilize two machine learning methods (Decision Tree and Linear Regression) to implement prediction model and obtain the preliminary results.**

***Index Terms*—Housing Price Prediction, Machine Learning, Decision Tree, Linear Regression**

## I. Dataset Description

We utilize American Housing Survey (AHS) 2017 as dataset [1]. Microdata are files containing individual responses to survey questions. They are used to create custom tabulations, allowing users to delve further into the rich detail collected in the American Housing Survey. In the AHS microdata, the basic unit is an individual housing unit. Each record shows most of the information associated with a specific housing unit or individual, except for data items that could be used to personally identify that housing unit or individual.

### A. Data Structure

- Household Table: Number of variables: 1089; Number of observations: 23084; Special cells: -9(Not reported), -6(Not applicable); Dtypes: float64(483), int64(70), object(536); Total size in memory: 152MB
- Mortgage Table: Number of variables: 21; Number of observations: 8859; Special cells: -9(Not reported), -6(Not applicable); Dtypes: float64(1), int64(3), object(17); Total size in memory: 789KB
- Person Table: Number of variables: 92; Number of observations: 49312; Special cells: -9(Not reported), -6(Not applicable); Dtypes: int64(15), object(77); Total size in memory: 18.1MB
- Project Table: Number of variables: 15; Number of observations: 22964; Special cells: -9(Not reported), -6(Not applicable); Dtypes: int64(1), object(14); Total size in memory: 1.52MB

### B. Where does the data set come from

The American Housing Survey (AHS) is sponsored by the Department of Housing and Urban Development (HUD) and conducted by the U.S. Census Bureau. The survey has been the most comprehensive national housing survey in the United States since its inception in 1973, providing current information on the size, composition, and quality of the nation's housing and measuring changes in our housing stock as it ages. The AHS is a longitudinal housing unit survey conducted biennially in odd-numbered years, with samples redrawn in 1985 and 2015 .

### C. What's the information in the data set

The survey provides up-to-date information about the quality and cost of housing in the United States and major metropolitan areas. The survey also includes questions about:

- the physical condition of homes and neighborhoods,
- the costs of financing and maintaining homes, and
- the characteristics of people who live in these homes.

Planners, policy makers, and community stakeholders use the results of the AHS to assess the housing needs of communities and the country. These statistics inform decisions that affect the housing opportunities for people of all income levels, ages, and racial and ethnic groups.

Since our country changes rapidly, policymakers in government and private organizations need current housing information to make decisions about programs that will affect people of all income levels, ages, and racial and ethnic groups.

## II. Problem Description

### A. Main Objective

According to the data given related to the house and house members, we will try to make prediction of:

- Houses' market values.
- Average salary of a house's living members.

We think that this is a very meaningful job since the model we trained could be used to predict a house's market value even before building it. Also we can predict the house members' average salary to give a reference to the advertisers.

### B. Difficulties We May Face

- The data set is large with ununified expression. The workload of preprocessing this data set is huge.
- We are trying to do the regression problems with thousands of features for each data point, the taining process will be both time and storage consuming.

## III. Methods and Experiment

This section describes the results of the project done so far. Firslty, we cost large amounts of time to pre-process dataset. Secondly, We utilize two machine learning algorithms to build housing value prediction model.

### A. Data Pre-processing

The original dataset includes 23084 rows and 3179 columns, which means the totally feature dimension is 23084*3179 . However, the original dataset includes much meaningless data and missing data (shown in figure1 and figure2). Therefore, the first step is to delete useless data and extract the key information from the original dataset as data feature.


Fig. 1. 37.7% data sample include missing value.

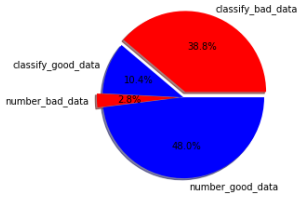


Fig. 2. 38.8% data sample include bad classify data and 2.8% data sample value are bad.

Based on the previous analysis and desired goal, we have divided data pre-processing step into following five steps:

*1) Delete data point whose market val is 'Nan':* 'market val' means the market value that is the prediction value. However, some market value are empty in dataset. These part cannot be involved in prediction model training so that these bad data sample will be removed.

*2) Delete columns with only one value:* When we check the dataset, we found that all of elements content are same in some columns. These data doesn't reflect any relationship with other data feature. Figure3 shows that 17.4% columns in dataset just have one same value. For this part, we delete them directly.

*3) Delete data features that has 25% missing values:* Some columns include too much empty to utilize in experiment and we cannot extract the useful information or build the correlation with other data sample. Thus, we directly these data columns. We set the threshold value as 25%.

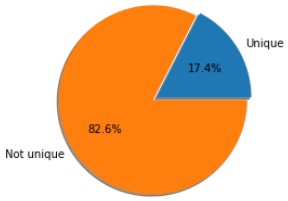


Fig. 3. 17.4% columns' value are unique.

*4) Change classification values to zeros and ones:* This step is similar with embedding operation. Because some feature are formed by string data type, these data value cannot be used as input or output. Therefore, we will embed these data value by using zero or one to represent.

*5) Using PCA to keep 90% data featuers:* After finishing previous four steps, we obtained the cleaned data feautre with 14373*713 size. It is still too complexity for prediction. Therefore, we apply Principal Components Analysis method to reduce the dimension of data feature [2]. Finally, the data feature dimension is reduced to 2D by using PCA (shown in figure4).

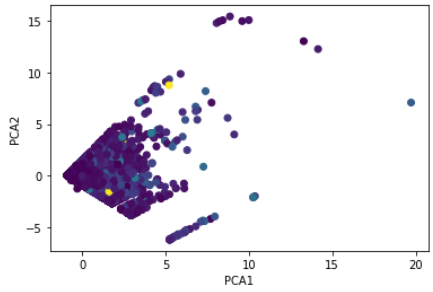


Fig. 4. Visualization for PCA result

Following the five operations as shown in above, the 2D data feature will be used to do regression prediction by using decision tree and linear regression. In addition, market value represent to the house price which are desired prediction value. Figure5 shows the value distribution for market value.

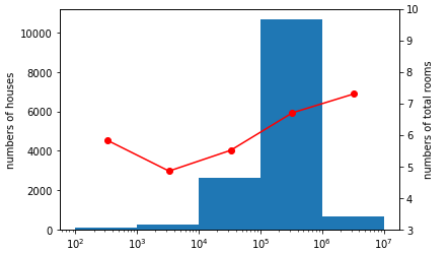


Fig. 5. The Distribution of Market Value

### B. Decision Tree Regression

Since our target functions has discrete real-valued outputs and the training data may contain errors and missing attribute values, we select decision tree model for our regression task [3]. By turning the parameters of the decision tree model, we could get different scores for the in-sample and out-of-sample. For example, by tuning the param-range, we could decide the best value range of the parameter that will be evaluated.


Fig. 6. Search the best parameter for decision tree regressor

From Fig 6, we know that with the rise of max depth of decision tree, the negative MSE value rises firstly, then drops because of overfitting. We can choose the best max depth size as 3. After retraining the model, we got the score of decision tree regression model for the data set (shown in TABEL 1).

### C. Linear Regression

$$\widehat{y}(\omega, \mathbf{x}) = \mathbf{b} + \omega_1 \cdot \mathbf{x_1} + \omega_2 \cdot \mathbf{x_2} + ... + \omega_p \cdot \mathbf{x_p} \quad (1)$$

Across the equation of the linear regression, we designate the vector $\omega$ as coefficient and $\mathbf{b}$ as intercept.

$$min_{\omega,\mathbf{b}} \sum_{i=1}^{n} (\mathbf{x}_i \cdot \omega + \mathbf{b} - \mathbf{y}_i)^2 \quad (2)$$

This is our optimization model using the least squares method. The goal of the linear regression is to predict the value of one or more continuous target variables $\widehat{y}$ given the value of a D-dimensional vector $\mathbf{x}$ of input variables, in such a way as to minimize the expected value of a suitably chosen loss function [4]. In our project, we fitted linear regression functions to data sets by minimizing the sum-of-squares error function. And this error function could be considered as the maximum likelihood solution under an assumed Gaussian noise model.

### D. Results Analysis

TABELE 1 summarize the experiment results. Decision tree can get the better result than linear regression. However, the ideal two scores should be larger than current results. Therefore, these two preliminary prediction results are not good enough. Firstly, may the dataset is not clean so that the irrelevant data features effects the prediction results. Secondly, the current model complexity is too simple to handle with this data features.

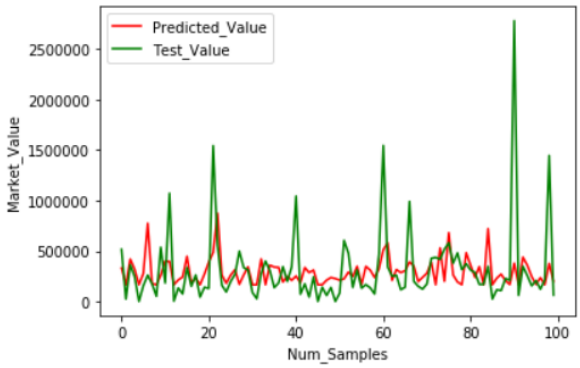


Fig. 7. Visualise the prediction results by using linear regression

TABLE I
SUMMARY OF THE EXPERIMENT RESULTS

| Model | Score for training data | Score for test data |
|---|---|---|
| Linear Regression | 0.124 | 0.122 |
| Decision Tree | 0.441 | 0.394 |

## IV. FUTURE WORK

From previous analysis, we found that the preliminary results are not so good. Therefore, we are trying to propose some improvement strategies to optimize the prediction results in the next half-stage. On the one hand, dataset should be deeply analyzed and cleaned. Because, the feature size is 14373*713 which still to large for the housing prediction and includes more irrelevant information. On the other hand, we research some advanced housing prices prediction algorithms and we will test the different methods [5,6]. For example, we can try to add penalty component in regressor and we can apply more algorithms to target this task. In addition, the current methods also should be continuous improved.

## REFERENCES


[1] American Housing Survey (AHS) 2017. https://www.census.gov/programs-surveys/ahs.html
[2] Dunteman G H. Principal components analysis[M]. Sage, 1989.
[3] Breiman L, Friedman J, Stone C J, et al. Classification and regression trees[M]. CRC press, 1984.
[4] Christopher M.Bishop, Michael Jordan, Jon Kleinberg, Bernhard S, et al. Pattern Recognition and Machine Learning Springer press, 2006.
[5] Lu S, Li Z, Qin Z, et al. A hybrid regression technique for house prices prediction[C]//2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE, 2017: 319-323.
[6] Fan C, Cui Z, Zhong X. House prices prediction with machine learning algorithms[C]//Proceedings of the 2018 10th International Conference on Machine Learning and Computing. 2018: 6-10.