CPE695 Final Project: Housing Prices Prediction on American Housing Survey 2017 Dataset

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

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

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

***Abstract*—Housing prices prediction is a significant indicator to reflect economic activities. Housing prices prediction aims to construct a stable mathematical prediction model by utilizing variant factors and housing prices prediction task can be seen as a regression problem. The project will focus on employ machine learning algorithms to build the prediction model. In this paper, we study and propose six machine learning algorithms to finish the task. Specifically, the best accuracy is 0.64 by using neural network in R2 evaluation standard. We record the all training detail and experiments results. In addition, we try deep learning based method and provide some potential improvement ideas.**

***Index Terms*—Housing Price Prediction, Machine Learning, American Housing Survey 2017**

I. Introduction

Housing prices prediction is a classical task in economic, finance and social science area. House prices is a indicator to reflect economic change and social development [1]. It is also a reference for a home buyer. More importantly, these predicted values can be utilized to many purpose, such as, second mortgages and home's insurance value [2]. Therefore, building a robust and efficiency house prices prediction model is necessary and challenge work.

In house prices prediction, this task can be formed as a regression problem. The previous work are focused on analyzing what factor will cause the house price change and create statistical model to predict prices. For example, empirical researchers often create hedonic price functions or hedonic models by using some factors [3]. Then, data engineering developing trigger researchers who start to construct more comprehensive data information. For instance, The Boston Housing Dataset contains 14 attributes collected by the U.S Census Service [4]. The abundant data features provides more aspects that can be used for analysis. In recently, inspired by machine learning technology developing, researchers take the more complex machine learning algorithms to deal with this task and achieved the better accurate [5].

However, these methods are still a lot of potential for improvement. Almost house prediction model just work on classical datasets in which it doesn't includes up to date data information. In this project, we propose to utilize American Housing Survey (AHS) 2017 dataset to make the housing price prediction. Figure1 shows the objective of the project. We utilize abundant data information as input and marketing value as label to build a strong prediction model. Firstly, we analysis and visualize AHS2017 dataset and then select useful data information. Secondly, we apply the Linear Regression, Decision Tree, Ensemble Learning methods (Random Forest, XGBoost and AdaBoost), Neural Network to AHS2017 dataset. Thirdly, we test different training strategy and data correlation analysis to optimize the model performance. Finally, we trying to study and implement deep learning based methods in this task.

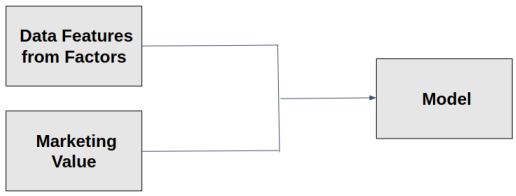

Fig. 1. The objective of the project

In this paper, we review the previous work about house prices prediction in section II. We introduce the detailed information about AHS2017 datasets in section III. Then, detailed experiment implementation step and experiment results are explained in section IV. We also do the analysis for the results and provide some possible improvement direction for future work in section V and section VI.

II. Related Work

Over the past 100 years, the relationship between the average annual growth rate of U.S. house prices and the inflation rate has accurately predicted U.S. house prices [6]. Whether investors' returns are high enough and whether they can achieve the purpose of diversifying their investment can be taken as a reference.As an investor, the questions are: are house prices expensive in the United States?What is the main pattern in the U.S. housing market?What is the rate of return on buying a home in the us now?

A. Inflation is used to predict house prices

According to the Casey Shiller (winner of the Nobel Prize in economics in 2013) index of house prices, the most authoritative one in the United States, from 1890 to 2013, house prices fell in 28 of the 123 years (23%) and rose in 95 (77%). The

deepest drop was in 2008, the worst year of the financial crisis, when it fell by 18%. Only twice in a row have they fallen for five years. The first time was during the great depression of 1929-33, when the decline was 26%. The second was the financial crisis caused by the bursting of the real estate bubble from 2006 to 2011, with a cumulative decline of 33%.

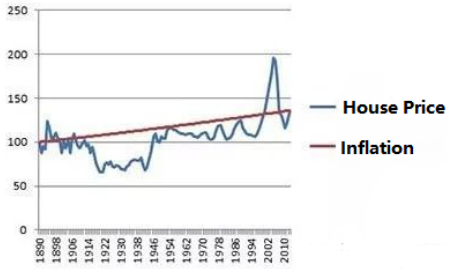

Fig. 2. US House Price and Inflation

We found that, after adjusting for inflation, house prices barely rose.This chart shows two patterns in America's housing market.First, the long-term growth rate of us house prices is 3.07 per cent, slightly higher than us inflation.Second, if house price changes deviate from this axis, no matter whether it is too high or too low, its subsequent trend can be predicted, that is, it will return to this axis.This axis also tells us that by 2013, after falling from 2006 to 2011 and rising from 2012 to 2013, house prices in the United States had returned to or even slightly above their historical average price after adjusting for inflation.This suggests that America's house prices will rise no higher than their historical average of 3% in the future.Other factors to consider include higher funding costs from the end of quantitative easing in the us and a rebound in the us economy.These two factors have opposite effects on house prices.

B. Rents are used to predict house prices

In addition to looking at inflation-adjusted house prices, another widely accepted measure of affordability and its direction is the ratio of rents to house prices.The logic is that rents represent real demand (not investment value), and the price of housing relative to that demand is a measure of how expensive or cheap it is.

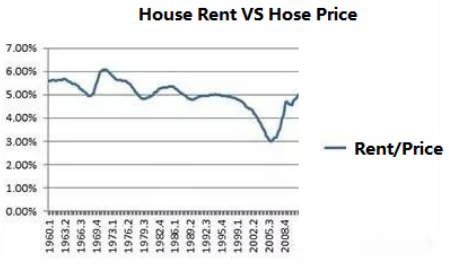

Fig. 3. US House Rent / House Price

As figure 3 shows, between the first quarter of 1960 and the first quarter of 2013, the average rent-to-house ratio in the United States was 4.98 percent.The real estate bubble in the early 21st century was manifested in a 38% decrease in the rent-to-house ratio from 4.8% in early 1999 to 2.98% in the first quarter of 2006.In 2007, three American economists (Davisa, Lehnertb, and Maninb) published a study of the overall rent-to-house ratio in the United States.Article now let a person to read Lin however, wrote: "according to the data since 1960, we built the rent to price ratio of time series, found that in 1960 to 1995 in the United States of the rent to price ratio between 5% and 5.5%, but soon fell after 1995, by the end of 2006, the rent to price ratio has reached a record low of 3.5%.If the rent-to-house ratio were to return to its historical average over the next five years, house prices would probably fall a lot."The subsequent story, as shown in figure 3, is that the rent-to-house ratio has risen substantially, approaching the historical average.The adjustment was made through falling house prices, including America's biggest single-year decline in more than 100 years in 2008.

C. Use historical prices to predict house prices

If home prices in Boston were discounted to $1 in early 1988 and rose to $1.08 in 2013, they would have grown at an annual rate of 0.3% after inflation, and experienced the same ups and downs as in New York.If the price of a home in Los Angeles had been discounted to $1 in early 1988, it would have risen to $1.12 by 2013, a rate of 0.46% a year after inflation.House prices in the United States have grown at an average annual rate of about 3 percent over the past 100 years, slightly higher than the nation's inflation rate (2.8 percent).If it rises much higher than inflation, there will be no growth, or a fall.The pattern is so precise that only 1943-1947 has been the only period of 100 years in which inflation has clearly outperformed and not fallen back.This rule applies not only to the United States, but also to first-tier cities.By 2013 the average house price in America had surpassed its historical average after inflation, not because it was cheap, but because it meant that the average annual rise in house prices in the future should be less than 3%.Whether this rate of return is high enough to achieve diversification is a matter of opinion.

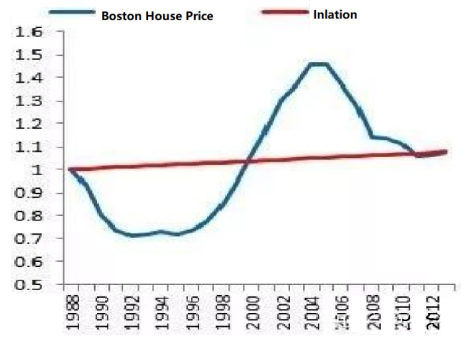

Fig. 4. Boston House Price and Inflation

D. Academic research on housing price forecasting

1) Modeling:

- Stevenson (2004) re-examined the issue of heteroscedasticity in hedonic house price models using Boston House Price Dataset.
- Fletcher, Mangan, and Raeburn (2004) argued that a wider range of diagnostic statistics should be used in the specification search for a good model, in particular, but not exclusively, those concerned with predictive stability.
- Bin (2004) estimated a hedonic price function using a semi-parametric regression and compared the price prediction performance with conventional parametric models.
- Bao and Wan (2004) illustrated how the technique of smoothing splines can be used to estimate hedonic housing price models.
- Kim and Park (2005) identified the spatial pattern of housing price changes and their determinants in Seoul and its neighboring new towns.
- Filho and Bin (2005) modeled a hedonic price function for housing as an additive non-parametric regression.
- Fan et al. (2006) utilized the decision tree approach, which is an important statistical pattern recognition tool in examining the relationship between house prices and housing characteristics.
- Kestens, Theriault, and Rosier (2006) introduced household-level data into hedonic models in order to measure the heterogeneity of implicit prices regarding household type, age, educational attainment, income, and the previous tenure status of the buyers.

2) Neural Network:

- Din et al. (2001) aimed to compare various real estate valuation models and the manner in which they take into account environmental variables.
- Kauko, Hooimeijer, and Hakfoort (2002) examined neural network modeling with an application to the housing market of Helsinki, Finland.
- Curry et al. (2002) viewed neural network modeling as a useful mean of specification testing, and hence their results imply some support for a linear formulation as an adequate approximation.
- Kauko (2003) evaluated the pros and cons of neural network models of property valuation in comparison with hedonic models, and provided some examples.
- Liu, Zhang, and Wu (2006) proposed a fuzzy neural network prediction model based on hedonic price theory to estimate the appropriate price level for a new real estate.

III. Dataset

We utilize America Housing Survey (AHS) 2017 as our data set [7]. 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;
- The characteristics of people who live in these homes, etc.

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.

IV. Experiment Design and Experiment Results

In this section, we describe the details of data preprocessing, and describe how to apply multiple machine learning regression algorithms. Then, we show the experiment results.

A. Pre-processing

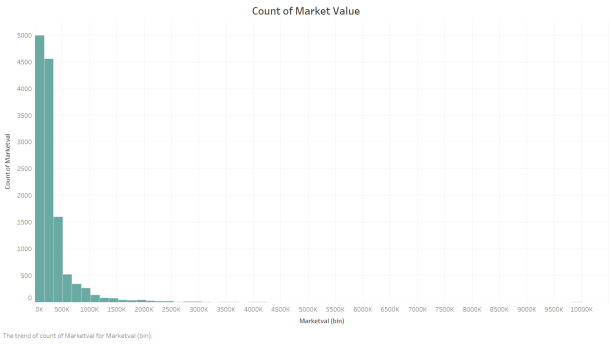

Fig. 5. Market Value Distribution

1) Step 1: Visulize the distribution of the Market Value: From figure 5, we know that the distribution of the market value is kind of like negative exponential distribution. Most of the market value gathers from range 0k to 1000k. Only few house price range from 400k to 1000k.This gives us an alert that high market value data may have a great influence on Linear Regression if we use Square Error as error function.

Fig. 6. Missing Market Value

2) Step 2: Dealing with Nan Value: From figure 6, we can see that there's nearly 37.7% data are missing market value. These data samples can't be used if the target value is missing.

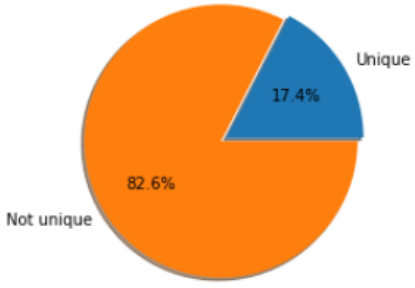

Fig. 7. Unique Features

3) Step 3: Feature Selection: From figure 7, we can see that there are about 17.4% of the features are unique features, which means that these features are the same with all the houses. So these features are totally useless for this regression problem.

We also found that there are a lot of features which have many Nan values. Although models like decision tree can deal with these Nan values as a specific labels, but up to 75% Nan values means that this feature didn't investigate properly. To this kind of features, we simply decided to delete them.

From figure 8, we can see that there are about 41.6% of the features contains at least 75% of the Nan values. Most of them are classification values.

Fig. 8. Good and Bad Features

Finally, we processed this data set into 14373 713. Then we begin to train different models using this data set. In the following paragraphs, we will name this data set as pre-processed data set.

B. Training Strategies

1) Simplest Regression: We firstly used linear regression and decision tree model to train our pre-processed data set (14373*713). For the linear regression model, we fitted LR's function to data set 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. Since our target function 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. Because of too many features, we only got awful results:

TABLE I
SUMMARY OF EXPERIMENT RESULT IN STRATEGY 1.

Model Name	R^2
Linear Regression	0.102
Decision Tree	0.312

In response to these awful results, we decide to use the PCA (Principal Components Analysis) method to reduce the dimension of the independent variables to get a better performance.

2) Use PCA to make dimensionality reduction: By taking the PCA method, we finally reduce the feature dimension to 2D (shown in figure 9).

So, we reduce our data set from 14373*713 to 14373*2 successfully. And the benefit of PCA on computation speed is pretty good. However, our prediction results are still awful:

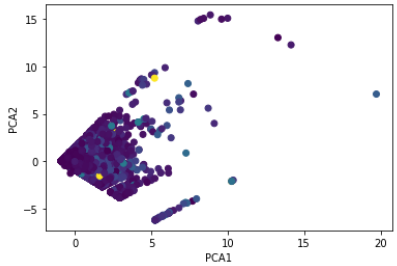

Fig. 9. Visualization for PCA result

TABLE II
SUMMARY OF EXPERIMENT RESULT IN STRATEGY 2.

Model Name	R^2
Linear Regression	0.122
Decision Tree	0.372

By analysing the result, we realize that the PCA could not get rid of those irrelevant variables such as House Holder's age, kids, insurance, etc. So it is necessary to implement a new strategy to screen out these irrelevant variables.

3) Artificially trim the features unrelated with the Market Value: We perform this strategy to remove irrelevant features by condition coefficients and we also drew a features correlation heat map (shown in figure 10).

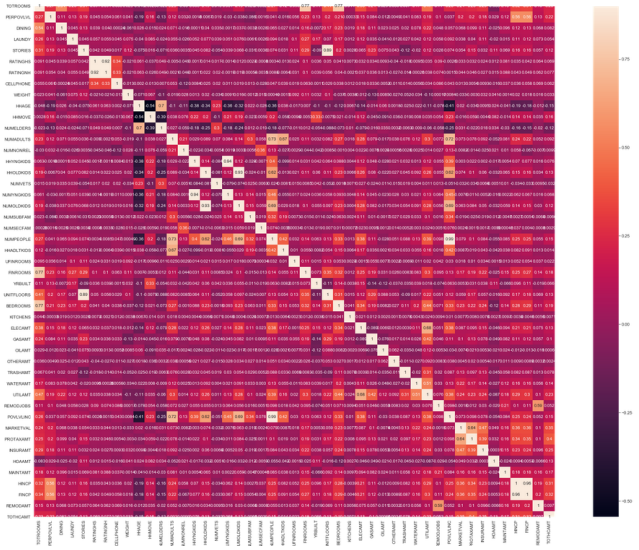

Fig. 10. Data Feature Correlation Analysis Heat Map

From the heat map, we could extract (e.g.Top 10) correlated features according to the color of the block or the correlation coefficients on the map. However, our models' scores are still not so good:

By observing our data set again, we inferred that those unsurveyed sample point might have bad influence on our model's prediction. Additionally, it is worth implementing some other machine learning algorithms to train our data set.

TABLE III
SUMMARY OF EXPERIMENT RESULT IN STRATEGY 3.

Model Name	R^2
Linear Regression	0.176
Decision Tree	0.394

4) Remove low investigated data and try more models: We found that many data have specific features that are not investigated. If a data sample's poverty situation is not investigated, then the other 23 features are also not investigated. We delete the data samples due to the poverty investigation. Then there's only 12902 data samples left.

We also tried more models. We added robust regression model since there are houses whose market price is not reasonable. But we didn't put these results here since they performs poorly. We also tried a lot of ensemble models based on tree models like Random Forest, Ada Boost, XG Boost.

TABLE IV
SUMMARY OF EXPERIMENT RESULT IN STRATEGY 4.

Model Name	R^2
Linear Regression	0.314
Decision Tree	0.407
Random Forest	0.403
Ada Boost	-0.22
XG Boost	0.412

From the result we can see that the performance is still need to be improved. Ada Boost had a very bad result bacause this model fits the bad performance data again and again. So the result may have a general balance to all points but not good in total result. So we thought that the extreme result may have a great influence on the final result. Our next method is to use unsupervised model to delete those outliers and see if the result gets better.

5) Using Isolation Forest to remove outliers: Isolation Forest is a quite new unsupervised learning technique invented resent years. The main idea is that it choose a random dimension and choose a random value from the highest to the lowest. So the data set will be cut into two parts. Store this data structure as binary tree. Continue cut the data set this way. So the outliers will be cut out faster than the clustered points statistically.

We cut the 10% outliers and do the same training to the rest data. This is the final result we got.

TABLE V
SUMMARY OF EXPERIMENT RESULT IN STRATEGY 5.

Model Name	R^2
Linear Regression	0.380
Decision Tree	0.472
Random Forest	0.462
Ada Boost	-0.311
XG Boost	0.217

From the result, we see that the improvement is so less. And the Ada Boost is getting even worse. We checked the remaining data samples and saw that there's still extreme

points inside there. The Isolation Forest didn't remove these points. But we still thinks that the extreme outliers are the key to the better performance. So we tried to remove the highest and lowest house market value by ourselves.

6) Remove extreme data manually: We tried to remove the first 10% and last 10% of the data manually due to the market price. Then train the data again. Here's the result.

TABLE VI
SUMMARY OF EXPERIMENT RESULT IN STRATEGY 6.

Model	R2 Score
Linear Regression	0.420
Decision Tree	0.407
Random Forest	0.471
AdaBoost	-0.22
XGBoost	0.412

We can see that the performance of the linear regression rose sharply, but other models remains the same. This is very reasonable because the remaining data samples' market values distributes as normal regression. And the extreme market values may contributes a very high Square Error in linear regression.

Fig. 11. Remaining data distribution

We still want to improve the performance of the regression. And the next method we tried is to change all the classification features into several features which is made of 0s and 1s. This may have an improvement.

7) Change classification features: Based on previous results, we are trying to make use of more data features to improve prediction performance. In removed part, we found that some factors have the useful meaning. However, these factors are categorical features and binary classification features. Therefore, we need to encode these data to match with input requirement. We transfer the data into 1 or 0 to represent and use one-hot embedding to process numerical that.

After processing categorical features, the input size become 10321 by 600. Then, we re-implement previous algorithms.

From above table, we see that some methods get the improvement. However, the random forest results became more worse. This result is reasonable. Firstly, we added more useful data as the feature so that the model can extract more important information to predict. Secondly, more data features are hard for random forest.

TABLE VII
SUMMARY OF EXPERIMENT RESULT IN STRATEGY 7.

Model	R2 Score
Linear Regression	0.492
Decision Tree	0.407
Random Forest	0.403
AdaBoost	-0.22
XGBoost	0.412

We still want to improve the performance of the regression. Therefore, we would like to propose some methods beyond the previous directions. Based on research, we implement neural network model to predict house prices.

8) Neural Network: Neural Network is a powerful methods to solve classification and regression problem [13]. Because neural network can build the complicated mathematical functions to process input data. Therefore, neural network can extract some implicit features and utilize these implicit content to do prediction. What's more, neural network will utilize loss function to compare the prediction results with true value and optimize parameter by using backpropogation algorithms. With the training iteration, the neural network maybe will find a local minimum point.

We implement a neural network by using Keras deep learning framework. Firstly, we split the whole dataset to 80% training part and 20% testing set. Then, Neural network has two hidden layer (fully connected layer) and each hidden layer has 64 hidden cells. For each fully connected layer, we add a RELU function as activation layer. In this experiment, Because this is a regression problem so that the output is one. Optimizer is rmsprop and loss function is MSE. We set batch size is 128 and train 1000 epochs.

TABLE VIII
RESULT FOR NEURAL NETWORK

Model	R2 Score
Neural Network	0.640

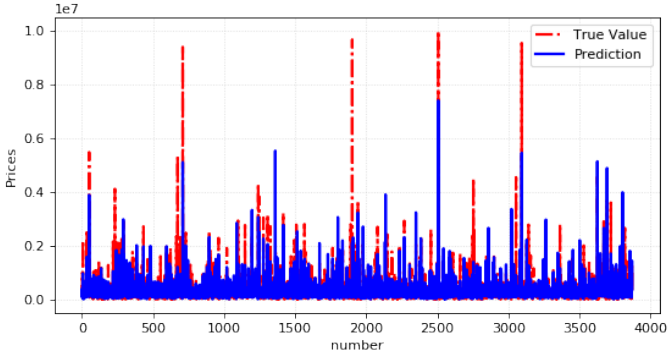

Fig. 12. Visualize Neural Network Prediction Result

Table 8 is result for neural network, neural work can achieve 0.64 R2 score, which is the best result in this project. The result make sense, because neural network may can find more deeply features and relationship between factors with market value. We visualize the regression results. Some points also have enough gap with true value. Therefore, we also can do more optimization in the furue work.

V. Results Analysis

We summary all final experiment result in Table9. Based on experiment results, we do some analysis for our training strategies and experiment design.

TABLE IX
SUMMARY OF PROJECT RESULTS.

Model	R2 Score
Linear Regression	0.492
Decision Tree	0.407
Random Forest	0.403
AdaBoost	-0.22
XGBoost	0.412
Neural Network	0.640

- There may still exist a better way to pre-process the data. But we really tried all methods we knew.
- There may exist better Deep Learning methods which performs better.
- Comparing to Boston House Price Dataset. We found that the features in this Data set in too detailed and irrelevant to the market price. Like position and city are two very important features which influences the market price. However, we don't have these features. Instead, we got many features relates to the house holder's privacy.

VI. Future Work

Based on experiment result analysis, we provides some potential improve ideas in this section.

a) Data processing: Out methods above is all about statistic analyzing. We didn't use our common sense to filter out the irrelevant data. Maybe after choosing features using our subjective judgement, the result will be better.

b) Deep Learning: In experiment, we also tried Convolutional Neural Network (1D conv layer) and Long-Short-Term-Memory [14], [15]. But the result is not good enough. We will try to analyze the performance and try more deep learning techniques in the future.

VII. Conclusions

Housing prices prediction is an important component in economic change analysis and other science areas. In this paper, we study and propose sit methods to solve this challenge, where these are Linear Regression, Decision Tree, Ensemble Learning methods (Random Forest, XGBoost and AdaBoost), Neural Network. We evaluate the proposed methods on American Housing Survey 2017 dataset, and the extensive experimental results show that the best R2 score is 0.640. In addition, we also tried deep learning based methods and we will do more optimization in the future.

References

[1] Case B, Clapp J, Dubin R, et al. Modeling spatial and temporal house price patterns: A comparison of four models[J]. The Journal of Real Estate Finance and Economics, 2004, 29(2): 167-191.

[2] Dubin R A. Predicting house prices using multiple listings data[J]. The Journal of Real Estate Finance and Economics, 1998, 17(1): 35-59.

[3] Ogwang T, Wang B. A hedonic price function for a northern BC community[J]. Social Indicators Research, 2003, 61(3): 285-296.

[4] Harrison Jr D, Rubinfeld D L. Hedonic housing prices and the demand for clean air[J]. 1978.

[5] Selim H. Determinants of house prices in Turkey: Hedonic regression versus artificial neural network[J]. Expert systems with Applications, 2009, 36(2): 2843-2852.

[6] Lancaster K J. A new approach to consumer theory[J]. Journal of political economy, 1966, 74(2): 132-157.

[7] American Housing Survey (AHS) 2017. https://www.census.gov/programs-surveys/ahs.html

[8] Dunteman G H. Principal components analysis[M]. Sage, 1989.

[9] Breiman L, Friedman J, Stone C J, et al. Classification and regression trees[M]. CRC press, 1984.

[10] Christopher M.Bishop, Michael Jordan, Jon Kleinberg, Bernhard S, et al. Pattern Recognition and Machine Learning Springer press, 2006.

[11] 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.

[12] 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.

[13] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[14] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[15] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.