STUDY THE TENDENCY OF HOUSE PRICE INDEX

A PREPRINT

Xitong Huo
Department of Computer Science and Statistic
University of Virginia
xh2kb@virginia.edu

Kaiwen Zhu
Department of Computer Science
University of Virginia
kz8pr@virginia.edu

Yongyi Li
Department of Computer Science
University of Virginia
yl9gq@virginia.edu

December 9, 2019

ABSTRACT

House prices, as if they are an eternal topic that plagues humanity. it has a Hugh impact on the development of society at all times. There was once an economist who had such an analysis of housing prices: 1. Rising house prices, generally encourage consumer spending and lead to higher economic growth. 2.sharp drop in house prices adversely affects consumer confidence, construction and leads to lower economic growth. 3. Rising house prices can also redistribute wealth within an economy – increasing the wealth of homeowners (primarily older people), but reducing effective living standards for those who do not own a house (often the young). And all this seems to indicate an incredibly powerful law. When we can grasp the law of housing price changes, we can approach the lifeline of economic change. We study the regression algorithms by using the house price index (HPI) in Virginia. Aiming at capturing the tendency of HPI with assistant sources, we plan to start from simple methods, including moving average, to complex methods, including regression tree and support vector machine etc. We have implemented some preliminary experiments and planed some future work based on it.

***K*eywords** · house price index · prediction · moving average · regression tree · support vector machine

1 Introduction

1.1 Motivation

A house price index (HPI) measures the price changes of residential housing as a percentage change from some specific start date (which has HPI of 100). Traditional statistics like Average Price and Median Sales Price fails to consider the effects of location, size, and public resources and so on. HPI, instead, is a more accurate index for the market.

In this project, we study the tendency of HPI, which can helps us to perform assessment and market analysis. Also, it can help us deepen the understanding of machine learning algorithms and data analysis techniques.

Specifically, we come up with some assumptions:

- HPI's tendency can be predicted within a relative error rate of 5%
- HPI is affected by the economic situation of the society, such as average salary, employ rate and so on
- Using these factors can improve the prediction of HPI.

In this paper, we will use simple methods, including moving average, to complex methods, including regression tree (and etc.) to validate our assumptions.

1.2 Dataset and Pre-processing

We have collected three datasets containing the HPI data from the website, they are AllTransHPI@0 [1], AllTransHPI@1 [2] and AllTransHPI@2 [3]. Before we start, we determine one from them.

AllTransHPI@0 contains HPI estimated from sales prices and appraisal data, using 1980:Q1 as a base of 100 units. The index is not seasonally adjusted. AllTransHPI@1 and AllTransHPI@2 both use 1995:Q1 as a base of 100 units. While AllTransHPI@2 contains two observation time stamp, resulting a slightly different in value.

We first check whether AllTransHPI@0 is consistent with AllTransHPI@1 using the same base. Unfortunately the answer is not. Also we have observed missing data in AllTransHPI@1. So, finally, we determine to use AllTransHPI@0.

Also we have collected two more datasets: YAURN [4], which records the unemployment rate of Virginia from 1976; and STTMINWGVA [5], which records the State Minimum Wage Rate for Virginia from 1976. We plot the three datasets in the same figure, and find out that: 1. As the minimum wage rises, HPI also rises, they are positive related; 2. there are some stable (or even decreasing) period of HPI, during when the unemployment rate is high, they are negative related.

All these datasets are from public government websites which means they are all reputable source. After simple data visualization and feature analysis, we found that there did not exist any missing value in our datasets. For future analysis and model construction, we created plot function for time-series structured data and made correlation analysis.

Figure 1: HPI v.s. Unemployment rate and State minimum wage

2 Methods

2.1 Moving average [1, 2]

A moving average (MA) is a widely used indicator in technical analysis that helps smooth out price action by filtering out the noise from random short-term price fluctuations. It is a trend-following, or lagging, indicator because it is based on past prices.

The two basic and commonly used moving averages are: the simple moving average (SMA), which is the simple average of a security over a defined number of time periods, $\hat{x}_{n+1} = \frac{1}{n}\sum_{i=1}^{n} x_i$, in which n is the window of time length; the exponential moving average (EMA), which gives greater weight to more recent prices, $\hat{x}_{n+1} = \alpha x_n + (1-\alpha)\hat{x}_{n-1}$.

[1] https://fred.stlouisfed.org/series/VASTHPI

[2] https://fred.stlouisfed.org/series/ATNHPIUS47260Q

[3] https://alfred.stlouisfed.org/series?seid=ATNHPIUS47260Q&utm_source=series_page&utm_medium=related_content&utm_term=related_resources&utm_campaign=alfred

[4] https://fred.stlouisfed.org/series/VAURN

[5] https://fred.stlouisfed.org/series/STTMINWGVA

2.2 Tree-based models

Regression tree [3] recursively split a node by seeking local-optimal variance reduction. The variance reduction of a node N is defined as the total reduction of the variance of the target variable x due to the split at this node: $I(N) = \frac{1}{|S|^2}\sum_{i\in S}\sum_{j\in S}\frac{1}{2}(x_i - x_j)^2 - [\frac{1}{|S_l|^2}\sum_{i\in Sv}\sum_{j\in S_l}\frac{1}{2}(x_i - x_j)^2 + \frac{1}{|S_r|^2}\sum_{i\in S_r}\sum_{j\in S_r}\frac{1}{2}(x_i - x_j)^2]$.

Regression tree can be boosted, forming gradient boosting tree [5]. When training, regression trees are added to fit the residual of prediction by the present models.

2.3 Support Vector Machine

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.[4] For here, we would like to use support vector regressor which is a regression method based on the theory of support vector machine.

3 Experiments

We use Root of Sum of Square Error (RMSE) for evaluation, by neglecting the start-up time series before a single time length window, **RMSE** can be calculated by $\sqrt{\frac{1}{T}\sum_{t=n}^{T}(x_t - \hat{x}_t)^2}$

Besides, we want to give some commercial advice. In this case, we expect to predict the up or down signal (s_t) of HPI accurately, which is helpful for commerce. We use $\mathbf{Acc} = \frac{1}{T-n-1}\sum_{t=n+1}^{T}\mathbf{1}_{\hat{s}_t=s_t}$ to evaluate our algorithms.

3.1 Moving average

In the first attempt, we used a relatively advanced statistical analysis method, time series(moving average). In statistics, a moving average (rolling average or running average) is a calculation to analyze data points by creating a series of averages of different subsets of the full data set. We do not find any libraries which contain the method to implement the moving average(it is a kind of time series method and you can find relative library in R), so we try to create the class which contains all functions we need.

We vary the time length window from 1 to 10, to predict the next time period HPI. Beyond our expectation, RMSE increases nearly linearly. This indicates **the recent HPI should be given greater weight**. However, the accuracy will not increase but decrease as the window length is too high, which prevent us simply taking the past HPI for present.

Table 1: RMSE v.s. window, simple moving average algorithm

window	RMSE	Acc
1	2.7125174613465726	0.7670454545454546
2	3.823685507892015	0.8
3	4.931868048089077	0.8218390804597702
4	6.056010587508376	0.8323699421965318
5	7.1946927292975476	0.8023255813953488
6	8.320208152696184	0.8245614035087719
7	9.425480401425096	0.8176470588235294
8	10.512310835861326	0.8165680473372781
9	11.57749668459072	0.8214285714285714

The result of EMA also reflects this insight. We fix the window size to be 4, with a discount ratio α varying from 0.6, 0.65, ... to 0.95. RMSE decreases nearly linearly, while accuracy also decreases.

In our real life, we would like to pay more attention to the tendency change of the HPI index instead of accurate number. For here, it is very impotent to achieve the balance between RMSE and tendency change of Accuracy.

Table 2: RMSE v.s. discount factor, exponential moving average algorithm

α	RMSE	Acc
0.55	6.459615063391246	0.7861271676300579
0.6	6.256673254610222	0.7803468208092486
0.65	6.066336495043807	0.7803468208092486
0.7	5.888446076187838	0.7630057803468208
0.75	5.722836511319793	0.7514450867052023
0.8	5.569338125682005	0.7456647398843931
0.85	5.427780229439994	0.7283236994219653
0.9	5.297994844932956	0.7283236994219653
0.95	5.179820922689625	0.7341040462427746

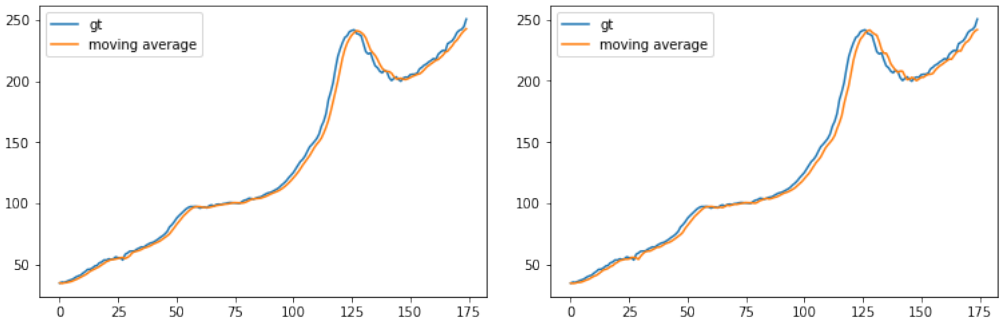

Figure 2: *left* SMA with window=3; *right* EMA with window=3

3.2 Tree-based models

Aforementioned, the previous HPIs are important features. As we want to include more features that reflect the financial situation. So here we use the unemployment rate and state minimum wage to be parts of input.

Note that here we use the history HPI with a time window of 4. This is because the accuracy to predict the tendency of window=4 is highest in moving average. It reflects our preference to knowing the tendency, which is more helpful for investment, but not just the value.

To avoid over-fitting, we restrict the maximum depth of regression tree to be 5. We randomly split the whole time series into training and testing sets, with a ratio 4:1.Better result can be obtained by using gradient boosting tree. We set the number of base regression trees in it to be 20. (n features=6, max depth=5, n estimators=20, regularization parameter=1.)

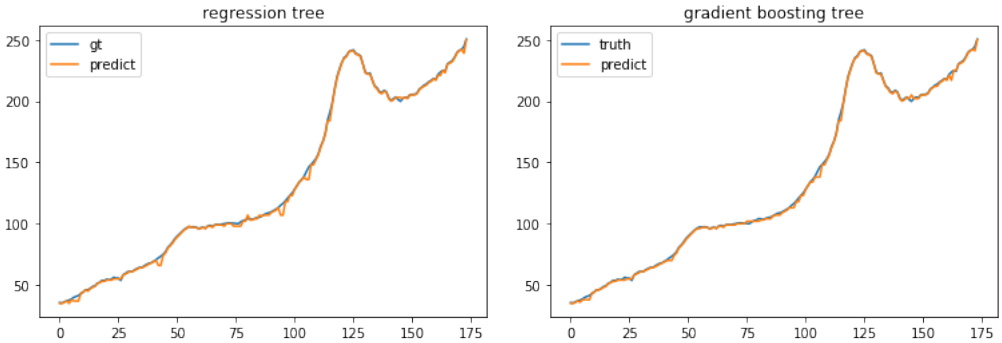

Figure 3: *left* regression tree; *right* gradient boosting tree

3.3 Support Vector Machine(SVR)

Based on the size of our dataset(174), it may be very efficient using SVM to predict the linear-unspreadable data. We created support vector regressor and a large parameter distribution grid as well(the size of dataset is very small, it would not spend long time tuning the hyper-parameters). To achieve the best result, we decided to use the gridsearch instead of random gridsearch.

There exists an interesting phenomenon from the construction process of SVR. When we tried to tuning the hyper-parameters of our model, the regularization term(C) had a huge impact on the final result, especially the RMSE. This may be because the time-series based data will fluctuate at a higher frequency in a very short time and these instabilities have caused serious interference in the fit of the model. The increased regularization term can reduce the effect from the instabilities, but it also contributed to the other problems, under-fitting. We tried to separate the datasets in different train-test split ratio(1:1). Like we thought before, there existed an exponential decrease in both RMSE and Acc, but on the whole, the model could still capture the relative changing tendency accurately.

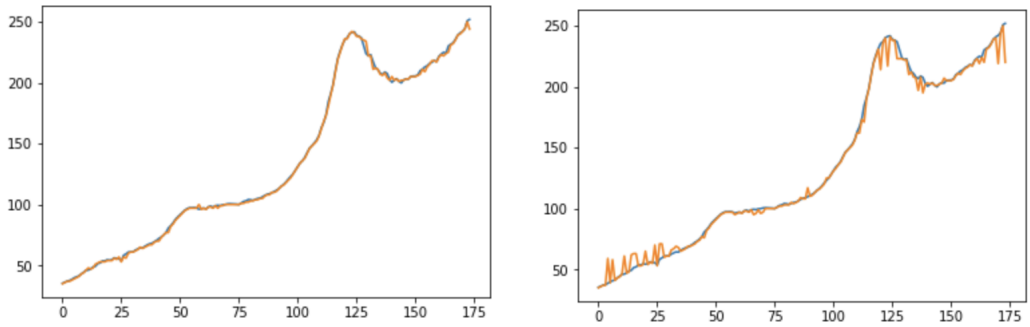

Figure 4: *left* SVR with C = 200; *right* SVR with C = 5

4 Results

4.1 Moving Average

The best result we get from our experiment is at window 4 and $\alpha = 0.55$ which is Acc = 0.8323699421965318 and 0.7861271676300579. Actually, there exists a RMSE and Accuracy trade off in our table. Although we usually use the RMSE as the standard principle to measure the efficiency of the model, the over fitting problems will also be extremely sever following the decrease of the RMSE.

4.2 Regression Tree

In this case, the regression tree's performance is much better than moving average, with RMSE = 3.905212098624755; tendency acc = 0.8554913294797688. We achieved better result in graident boosting regression tree method. With ensemble learning, we boost the performance again, with RMSE = 2.794064246852231 ; tendency acc = 0.86705202312138 (n features=6, max depth=5, n estimators=20, regularization parameter=1.) Also, compared with rest of our methods, the tree-based model achieved the best result and there is no doubt that we will use such model to predict the HPI in the future.

4.3 SVR

The best result we got from SVR is RMSE = 2.8308191466883668; tendency acc = 0.838150289017341 with hyper-parameters (C=500, cachesize=200, coef0=0.0, degree=3, epsilon=0.1, gamma=0.05,kernel='rbf', maxiter=-1, shrinking=True, tol=0.001, verbose=False).

5 Conclusion

To this end, we have verified the guess before we implement codes and analyze the results. We come out with some conclusions:

1. HPI is predictable: From the traditional statistical moving average method, to the advanced machine learning method regression and SVR, there is no doubt that the HPI index can be predicted. For existing data, the model shows a very good fit, especially the regression which achieve the best results in both RMSE and tendency accuracy(2.794064246852231 and 0.86705202312138). With the higher prediction accuracy of HPI, the residents in Virginia could know more about the value of the house they care about and how to manage their wealth wisely.

2. HPI is affected by financial situation: After we add more collected relative data the model shows better performance, and the previous data pre-processing also shows the correlation between HPI and economic conditions. Instead,the Virginia economists could use HPI to determine the economic conditions of Virginia, and help the government to step into introduce appropriate policies and counter-measures to address the problem. We will consider more details of financial situation and collect more data from websites.

3. Compared with traditional statistical analysis methods, machine learning models show better prediction results. This approach is considerably more convenient and more significant for Virginia's residents than using traditional methods. People could fully describe to get a exacting evaluations of the house, since we had more categories.

However, there still exist some questions after finishing the whole project process. There are strong concerns over the accuracy of technology and its impact on privacy. For the future, we will try to improve the prediction accuracy, and find more room for enhancement in data pre-processing, such as eliminating errors, noised, in-consistent data or missing data that are contained in the data set. In the meantime, we would determine if the application of deep neural networks have a better effect on the processing of non-structural data, and determine the models we created were either under-fitting or over-fitting. Machine learning is always full of knowledge waiting for your exploration. Machine learning will Never stop, this is what we learned from this project.

6 Contributions

Overall, each member in our group contributed equally and fairly to the group. Some of the main responsibilities of each member are given below. Note that this is not a complete description as every member worked with pretty much every part of the project. Yongyi Li focused on data preprocessing, which includes the basic data visualization and basic feature engineering. He also makes a contribution to SMA and edited the first part of the paper. Kaiwen Zhu focused on the EMA, making the basic model of regression trees (default parameters and main class), SVR, and editing the rest of the paper. At last, Xitong Huo moved to optimize the regression tree model by using grid search and random search to find the best combination of the model, and feature collecting. And all of us did the overall checking together.

References

[1] https://en.wikipedia.org/wiki/Moving_average

[2] https://en.wikipedia.org/wiki/Regression_analysis

[3] Loh, Wei-Yin. "Classification and regression trees." Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1 (2011): 14-23.

[4] https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[5] Ganjisaffar, Yasser, Rich Caruana, and Cristina Videira Lopes. "Bagging gradient-boosted trees for high precision, low variance ranking models." Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011.