Real Estate Price Modelling

Long Term Project

Data Science in Quantitative Finance
Fall 2018

Siyue Su & ss11378@nyu.edu
Sulin Liu & sl6501@nyu.edu
Bixing Yan & by783@nyu.edu

Nov 28th, 2018

1. Executive summary

The main objective of this project is to discover what drives real estate price at nation, state and county level. This could help us to price our homes and predict the future movement of real estate. Our approach is to identify covariance features and alpha features for real estate price return by nation, state and county level data mining and data analysis.

2. Motivation and problem formulation

- **Background and Motivation**

Zillow already has published a basic 12-month home value forecasting online at the State, County, Zip code and Metro level. However, while their model uses market dynamics data such as price momentum and other techniques. Zillow's forecasting model does not fully justify their choice of features. In addition, it skips a lot of details and sometimes refer too much to economic intuitions. While it is a good idea to read these references and source inspiration from econometric models of others, an alternative approach will be to train a properly constructed "kitchen sink" model and let the data itself tell you what the relevant features are themselves.

- **Question Raised**

We want to examine what are the main drivers for median real estate prices return at national, state, county and metro/zip-code levels, and figure out the important features rather than citing too much intuitive insights from existing economic models.

3. Dataset

We gather datasets from 3 different levels: national, state/metro city and county. We will briefly describe the statistical properties of each dataset.

- **National Long-term Data**

1. **Nation level house prices dataset, US Home Prices 1890-Present**

   The data is downloaded from http://www.econ.yale.edu/~shiller/data/Fig3-1.xls.

   This national level home prices dataset describe the national level economic condition of the US house since 1890. Besides the information about the house itself like the home price index and home cost index, many other macro level economic quantity are also available in this dataset, like the US population, interest rate and CPI. Thus, this is our main data source for our national level house price model.

2. **The data is downloaded from**

   http://www.econ.yale.edu/~shiller/data/ie_data.xls and
   http://www.econ.yale.edu/~shiller/data/chapt26.xlsx

   Beside the macro economic condition, the condition of stock market also have a very significant influence on the house price. When the stock market performs strong, it will absorb a large amount of money, from the view of investment purpose, which many take up the capital available curre for real estate market. However, on the other hand, the strong performance in stock market implies a prosperity of future economy, from the view of living purpose, which many strengthen the confidence of ordinary customer and prompt them to purchase the house. Thus, the influence of stock market on the house price should be discuss case by case, Here, we will discuss it with data analytics.

- **State and Metro Data**

1. **State level historical house prices on Quandl gathered by FMHPI**

   The data is downloaded from https://www.quandl.com/data/FMAC/HPI-House-Price-Index-All-States-and-US-NationalThe FMHPI provides a measure of typical price inflation for houses within the U.S. Values in the table are calculated monthly but are released at the end of the following quarter. All series consists monthly data that begins in January 1975 and ends in December 2017. The national index is defined as a weighted average of the 50 state indexes and Washington, DC. The FMHPI is based on an ever-expanding database of loans purchased by either Freddie Mac or Fannie Mae. The date also includes seasonalized and un-seasonalized price aggregated for each month.

2. **Case-Shiller indices of 20 Metro cities**

   The data is downloaded from https://fred.stlouisfed.org/series/SPCS20RSA. We examine the Case-Schiller index methodology from https://us.spindices.com/documents/methodologies/methodology-sp-corelogic-cs-home-price-indices.pdf and find out the component metropolitan areas of its S&P/Case-Shiller 20-City Composite Home Price Index. We then download data (Seasonally Adjusted) of those cities and merge them together. These 20 metro cities include Atlanta, Boston, Cleveland, Chicago, Charlotte, Dallas, Detroit, Denver, Las Vegas, Los Angeles, Miami, Minneapolis, New York, Phoenix, Portland, San Diego, Seattle, San Francisco, Tampa and Washington DC. In addition we manually scrape down the data from https://www.nber.org/papers/w2393.pdf, and merge it with the previous Case-

Shiller full table, so that we can look at the effects of the 1970s inflation on real estate for San Francisco, Dallas, Chicago and Atlanta.

3. **Economic and Demographic Data**
    1) **Residence Population:** State-level annual data from 1970 to 2018.
    2) **Crime Dataframe:** State-level crime statistics including violent crime total/rate, property crime total/rate, Robbery total/rate, Assault total/rate and Rape total/rate etc. It is downloaded from https://www.ucrdatatool.gov/. The data set ranges from 1960 to 2018 documented yearly.
    3) **Homeownership Rate:** State-level homeownership rate annual data from 1970 to 2018, adjusted by state total population.
    4) **All-transaction House Price Index:** State-level AHPI annual index from 1975 to 2018. Since the index is lagged and involve information into the future, we neglect it in our alpha models.
    5) **State Minimum Wage:** State-level annual data from 1968 to 2018. In this dataset it contains many missing values. We mainly use EM-lite linear regression method to fill the missing data.
    6) **Real GDP:** State-level annual data from 1980 to 2018.
    7) **Median Household Income:** The dataset is scrapped from https://www.irs.gov/statistics/soi-tax-stats-products-publications-and-papers, and contains various tax data for each state. The dataset ranges from 1980 to 2017.
    8) **Wage & Salaries:** The dataset is also scrapped from https://www.irs.gov/statistics/soi-tax-stats-products-publications-and-papers, and ranges from 1984 to 2017.
    9) **Target y-variable Zillow state level real estate price prediction:** The dataset is obtained from https://www.zillow.com/research/data/. It includes Zillow's prediction and should be compared with our state-level real estate price prediction.

Note that all the above datasets except the last two are scrapped from https://fred.stlouisfed.org/ or https://www.census.gov/programs-surveys/popest/data/data-sets.html. In this case, the census data is not annual, and thus we resample it so that it is transformed into annual data in order to make sure it is synchronized and that we are not looking into the future. Then we merge it with all the remaining datasets to obtain the full state-level data.

- **County Data**

What we want to predict is ZHVI index for counties:

ZHVI data: Zillow-defined property price index. It's a representative for propriety value. We use the median ZHVI per sqft for each county as the independent variable.

EDA data analysis:

*1. **Crime data:** we use 'violent crime total' in dataset provided by uniform crime reporting program in U.S department of Justice*

2. ***Education data:*** We have education score data for each county from United States census website:
https://www.census.gov/support/USACdataDownloads.html#EDU

3. ***Income data:*** We have education score data for each county from United States census website:
https://www.census.gov/support/USACdataDownloads.html#INC

Factor engineering and prediction data:

Data are from zillow's research:

1. ***Increase data:*** percentage of houses whose value increase last year

2. ***Decrease data****:* percentage of houses whose value decrease last year

3. ***Sales data:*** Median of sales price

4. ***List price data:*** Median of listed price

5. ***Affordability data:*** Zillow quarterly affordability indices, that evaluate mortgage affordability in given areas, it's an index calculated by mortgage payment for the median-valued home, median household income, etc. see reference: https://www.zillow.com/research/data/

6. ***Mortgage rate:*** is the average mortgage rate quoted on Zillow for a 30-year, fixed-rate mortgage in 15-minute increments during business hours, 6:00 AM to 5:00 PM Pacific. It does not include quotes for jumbo loans, FHA loans, VA loans, loans with mortgage insurance or quotes to consumers with credit scores below 720. Federal holidays are excluded.

7. ***ZHVI forecast****:* The ZHVF is the one-year forecast of the ZHVI, see reference:
https://www.zillow.com/research/2013/01/24/zillow-home-value-forecast-methodology-2/

4. Methodology

There are two main objectives for our project: (1) Calibrate ZHVI index to Case-Shiller index (2) Discover real estate factors, build and train the models. We will address our approaches to these two problems respectively:

I. Calibrate ZHVI index to Case-Shiller index

We first try to calibrate Case-Shiller index to ZHVI by transforming Case-Shiller in some specific ways, and then to see if one the them monitors the other well and spot any difference in those two indices. Note that both indices are on the metro city level, which includes 20 major cities in the U.S. The time series for both range from 1996.5 to 2018.10 for most of the cities and are quoted monthly. Then we try to build a model to see if we could successfully predict the value of one of the index using the lagged value of the other index.

II. **Discover real estate factors, build and train the models**

We divide the problem into three levels: nation level, state level and county level. Our team members would research on factors separately in these levels. Finally we could discuss whether drivers of price return would vary by levels. For each level, we generally follow the these steps:

1. Data gathering and processing
2. Factor construction
3. Identify risk factors using various models and machine learning techniques (OLS as benchmark)
4. Identify alpha factors using various models and machine learning techniques (OLS as benchmark)
5. Make predictions based on risk factors and alpha factors we found
6. Compare our prediction to Zillow forecast models

5. Results

Here are some exploratory data analysis results:

ZHVI Calibration to Case-Shiller

1. Calibration and Spread Calculation

ZHVI index for metro cities is Zillow's own prediction for the house price. Since ZHVI Index methodology removes foreclosures from their estimates and since ZHVI measures the median house price, their data should be inherently upward-biased relative to the Case-Shiller data. Therefore the first thing we want to do is to transform the two index for calibration purpose. For Case-Shiller data, we divide the Case-Shiller index by the corresponding ZHVI index. Then we normalize both indices to be 1 on the first available data for ZHVI by dividing the index by the very first value for each metro city. The calibration result is shown below.

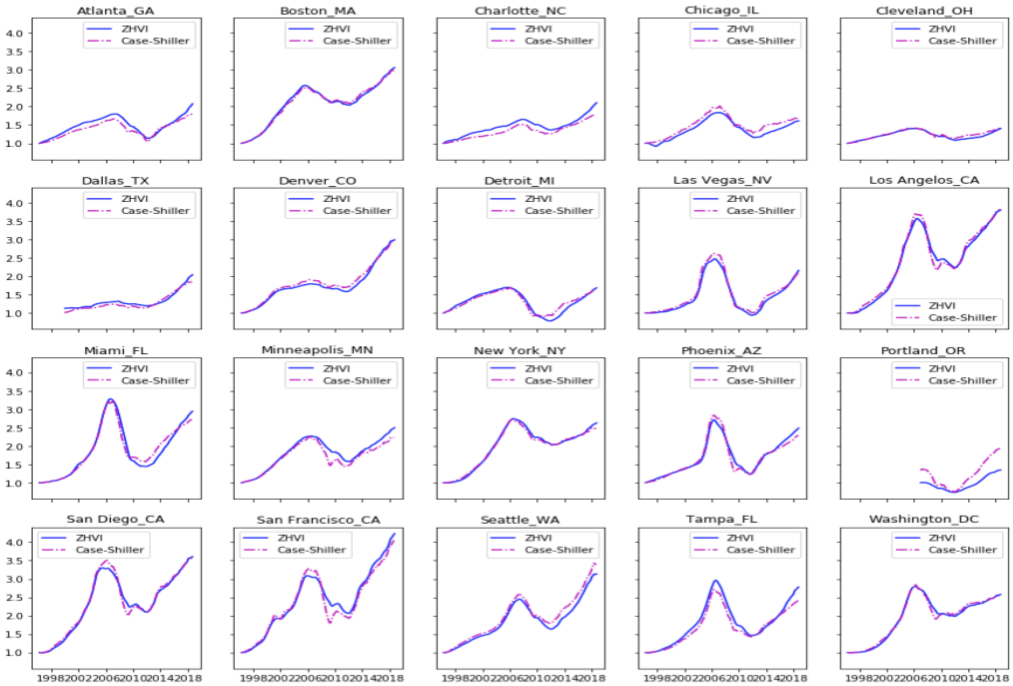


Fig: ZHVI Calibration to Case-Shiller for all metro cities

From the above plot, we find out that actually not all the ZHVIs are upward biased. The reason for the upward biased ZHVI might due to the fact that the house price for that metro city has fat tails, and thus making its median value upward biased. Overall the two indices overlap with each other and tend to have the same up and down trend over time. One thing to notice is that for Portland, OR, the two indices does not follows each other very well. The main reason is that Portland, OR has the fewest prediction data.

Next we go ahead and calculate the conversion factor between the two indices based on their normalized values to check if the resulting multiplicative factors will allow us to translate statements about ZHVI to statements about Case-Shiller. Spread factor is also calculated to see if this reveals information about foreclosures.

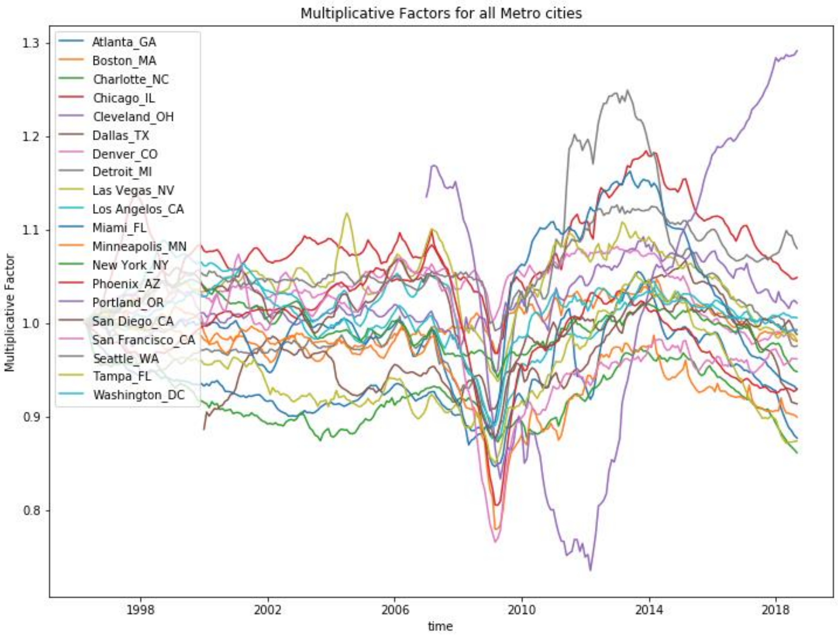
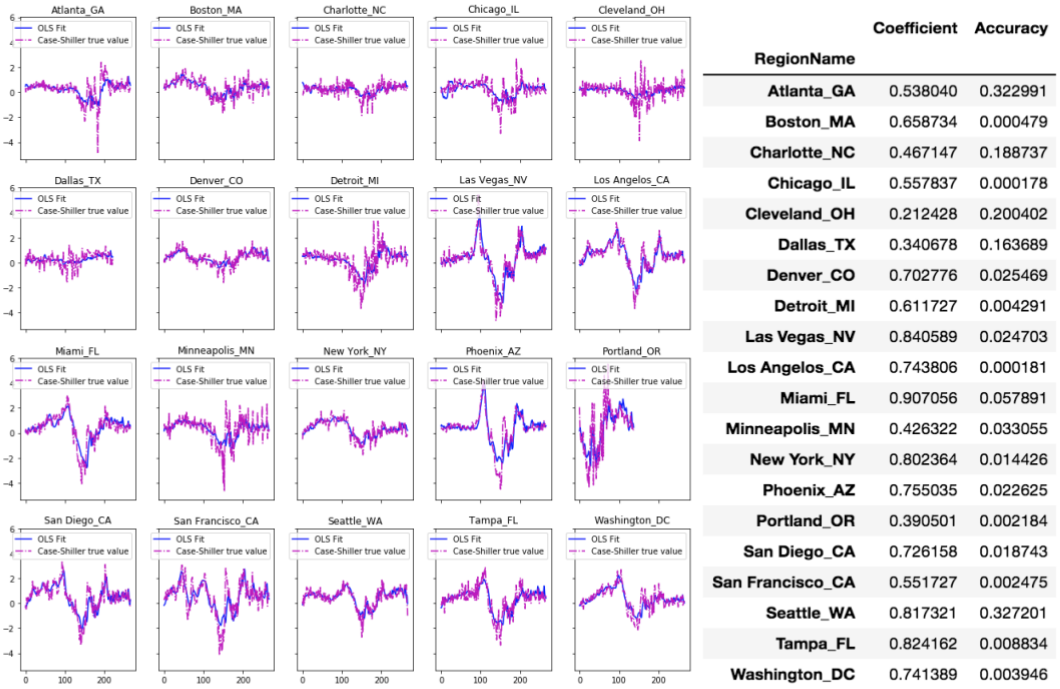


Fig: Conversion factor for all metro cities

If the conversion factor is below 1, it means that the difference between the two indices is negative, which means that ZHVI is upward biased. From the plot we can see that during 2009 ZHVI is upward biased for most metro cities due to high percentage of foreclosures.

2. Build Prediction Model

Since in the previous section we see that the two indices tracks each other very closely, we want to see if we can predict the C-S index returns at time t from the ZHVI index returns ones at time t-1, or vice versa. Therefore we build OLS model to do the prediction for Case-Shiller using ZHVI as regressors. The model is built using the percentage change of the two indices. Here is the result of the model.



RegionName	Coefficient	Accuracy
Atlanta_GA	0.538040	0.322991
Boston_MA	0.658734	0.000479
Charlotte_NC	0.467147	0.188737
Chicago_IL	0.557837	0.000178
Cleveland_OH	0.212428	0.200402
Dallas_TX	0.340678	0.163689
Denver_CO	0.702776	0.025469
Detroit_MI	0.611727	0.004291
Las Vegas_NV	0.840589	0.024703
Los Angelos_CA	0.743806	0.000181
Miami_FL	0.907056	0.057891
Minneapolis_MN	0.426322	0.033055
New York_NY	0.802364	0.014426
Phoenix_AZ	0.755035	0.022625
Portland_OR	0.390501	0.002184
San Diego_CA	0.726158	0.018743
San Francisco_CA	0.551727	0.002475
Seattle_WA	0.817321	0.327201
Tampa_FL	0.824162	0.008834
Washington_DC	0.741389	0.003946

Fig: OLS fit of Case-Shiller index & Results from OLS

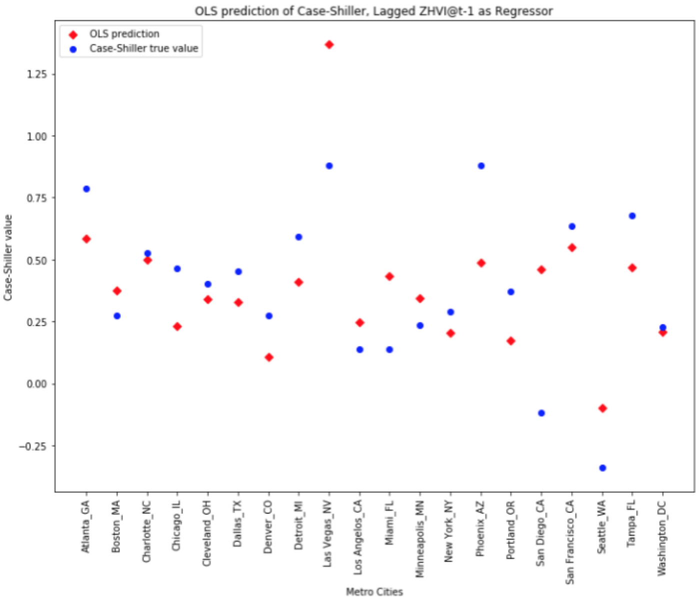
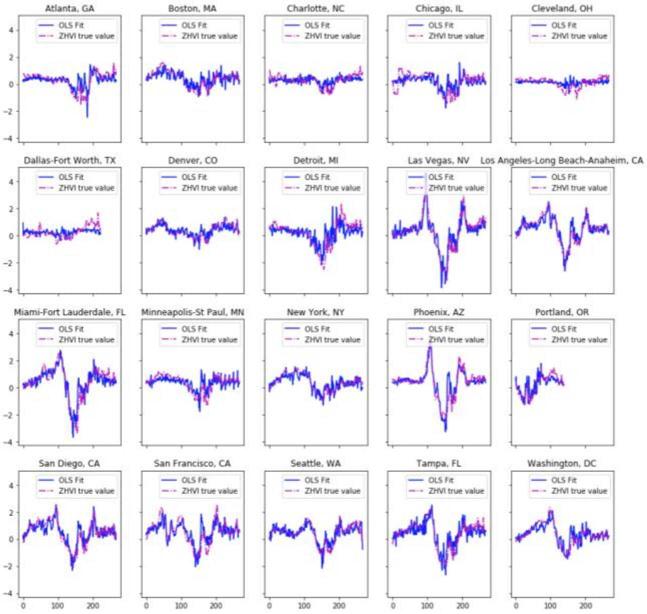
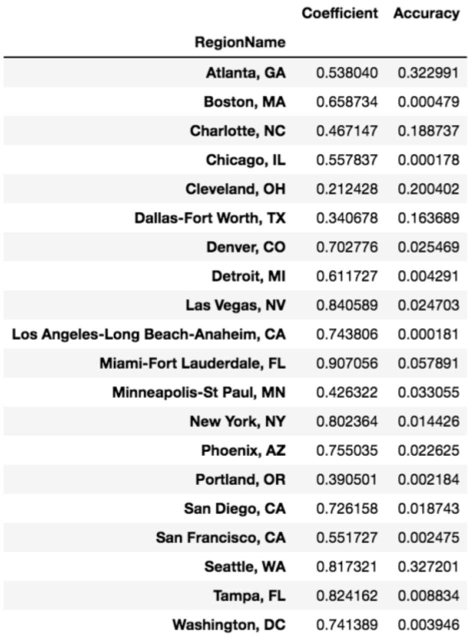


Fig: OLS prediction for Case-Shiller for all metro cities

From the coefficients of OLS, we can see that for most state, the coefficients are very high and all exceeds 0.5 in absolute value, indicating that lagged ZHVI value is a good candidate for prediction Case-Shiller value. From the prediction we can see that the prediction is indeed very good, with the corresponding accuracy score listed in the table. We measure the accuracy using mean squared error.

Then we do the other way around and find out the Case-Shiller is not a good leading indicator for ZHVI. From the results below, we can see that the coefficients are not very significant.





RegionName	Coefficient	Accuracy
Atlanta, GA	0.538040	0.322991
Boston, MA	0.658734	0.000479
Charlotte, NC	0.467147	0.188737
Chicago, IL	0.557837	0.000178
Cleveland, OH	0.212428	0.200402
Dallas-Fort Worth, TX	0.340678	0.163689
Denver, CO	0.702776	0.025469
Detroit, MI	0.611727	0.004291
Las Vegas, NV	0.840589	0.024703
Los Angeles-Long Beach-Anaheim, CA	0.743806	0.000181
Miami-Fort Lauderdale, FL	0.907056	0.057891
Minneapolis-St Paul, MN	0.426322	0.033055
New York, NY	0.802364	0.014426
Phoenix, AZ	0.755035	0.022625
Portland, OR	0.390501	0.002184
San Diego, CA	0.726158	0.018743
San Francisco, CA	0.551727	0.002475
Seattle, WA	0.817321	0.327201
Tampa, FL	0.824162	0.008834
Washington, DC	0.741389	0.003946

Fig: OLS fit of ZHVI index & Results from OLS

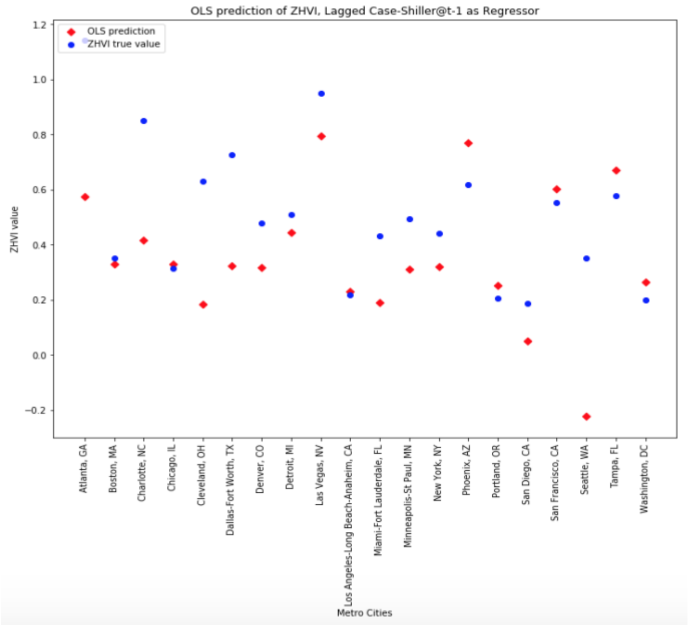


Fig: OLS prediction for ZHVI for all metro cities

Nation level:

1. **Overview national level house price dataset**

Our purpose is to find the relation between national level house price and macro economic factors. Thus, after removing some redundant information like information source, we first plot the price and relevant quantities are shown as following:

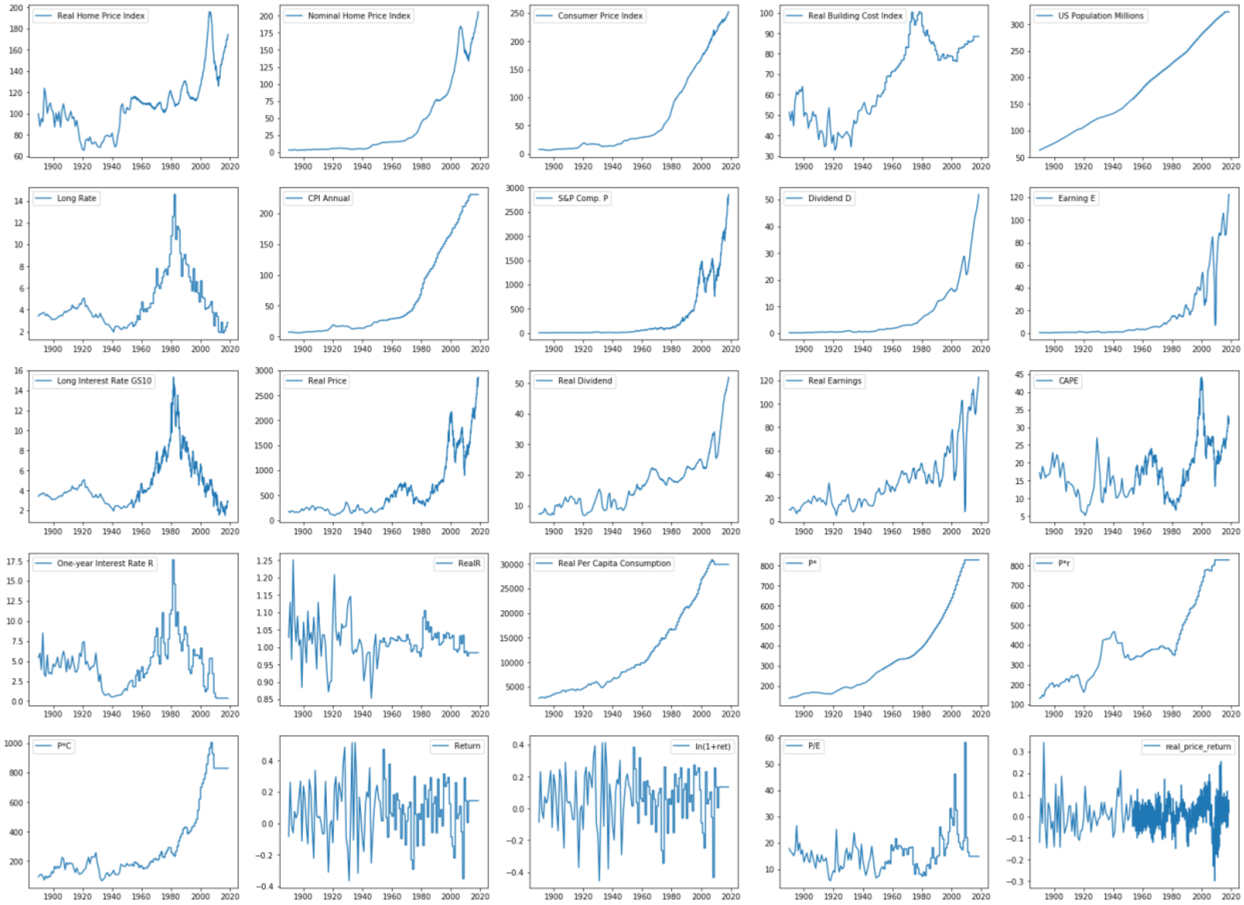


Fig. 1. Plot of Home Price Index and Related Economic Parameters.

From the above figure, we can see that, there two different kinds of home price index, real and nominal. So one question here is which index can better reveal the relation between home price and macro economic factors. Moreover, in most time series problems, the quantities we predict is not price, but the return. So the first problem for us is to choose a our prediction target.

In terms of relation between different quantities, as its name, correlation is always a good metric to start with. Thus, we first plot the correlation matrix of the price index and other economic parameters, as shown below:

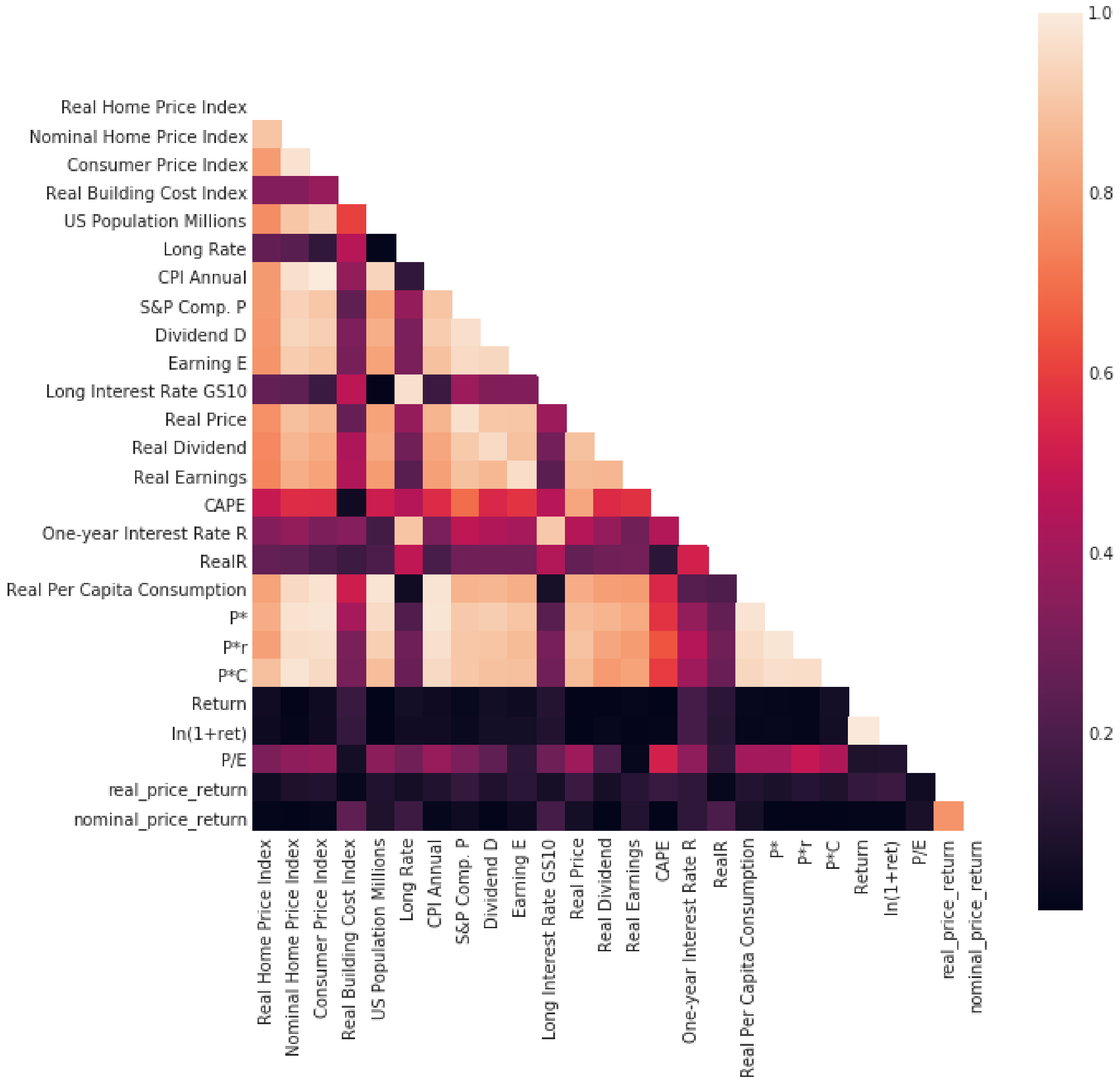


Fig. 2. Correlation Matrix of Home Price and Related Economic Parameters

From the above figure, we can find that, compared to the nominal house price, the real house price index have a higher correlation with other quantities. Thus, here we choose the real house price index as our target value. In terms of the return, both the real and nominal return have a very low correlation with other economic parameters. This may be because the return is first order difference of the price and may have better correlation with difference of other economic factors. However, this requires a lot feature engineering process and thus we will not use the return to keep our model more succinct and intuitive.

2. Modeling and Evaluation

Since our purpose is to find the underlying relation between home price and macro economic factors. The interpretability is the first priority. Thus, we plan to use tree algorithms here. Among different tree models, the boost tree usually give good result. Thus, here we would use Xgboost as our model. The prediction of the Xgboost model is shown as below:

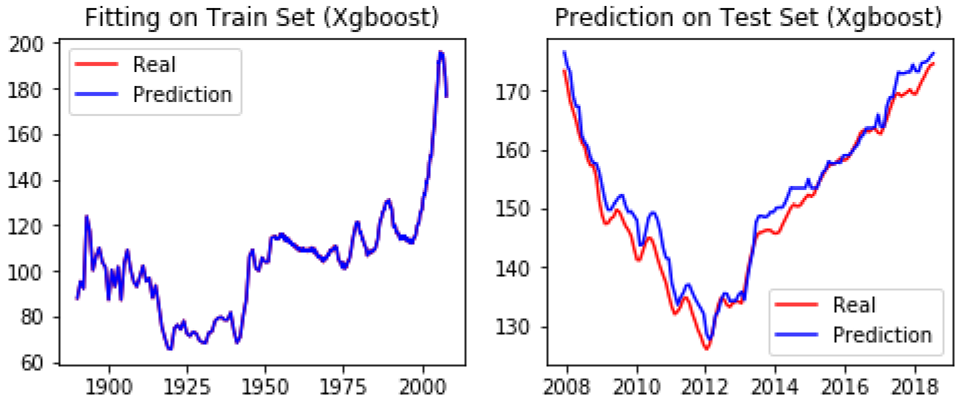


Fig. 3. Performance of Xgboost on Train Set and Test Set

From the above figure, we can see that, the model fairly predict the characteristics of the price. The price trend are well forecasted, while the standard deviation on the prediction of the test set is 2.86 (5% compared to the amplitude of the price) which is also an acceptable result. This indicates that our fitting is successful.

The importance (f score) of different features are shown in the form below

Feature	Long Rate GS10	S&P Comp. P	CAPE	CPI	Real D	Real P	Real E	E	Real Building Cost	Real R	D	R	Long Rate	P/E	P*C	P*r	Population	P*	Real PCC
f score	443	348	344	321	311	309	308	242	156	145	126	94	84	78	72	58	52	20	13

Form. 1. F Score of Different Macro Economic Parameters

3. Discussion

Although we have gotten a very good result and figure out the importance of given features, we doubt that can we still improve the accuracy or whether there is some other important hidden information?

Thus, tentatively we used a classical model, ARIMA(p,i,q), to predict the home price.

The real home price index and its autocorrelation and partial autocorrelation are shown below:

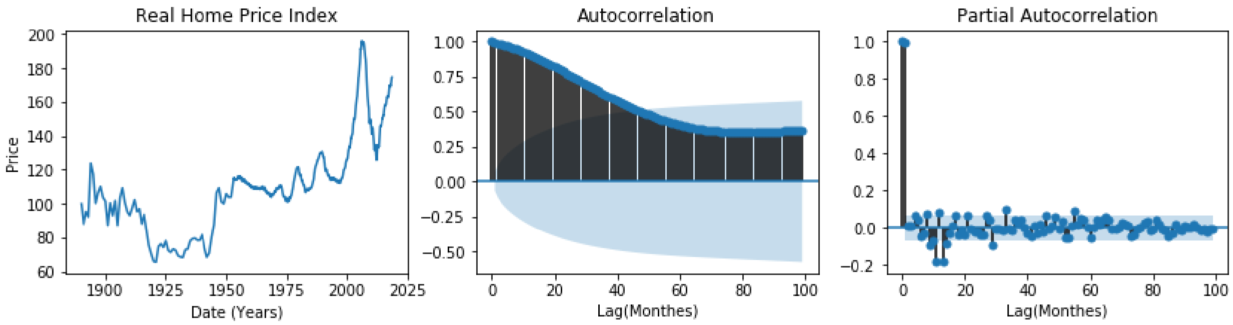


Fig. 4. Price and its Autocorrelation and Partial Autocorrelation

From the above figure, we can know that, the real home price cannot be described with ARMA process, i.e. the integral parameter i is not 0. So we plot the difference real home price index and its autocorrelation and partial autocorrelation are shown below:

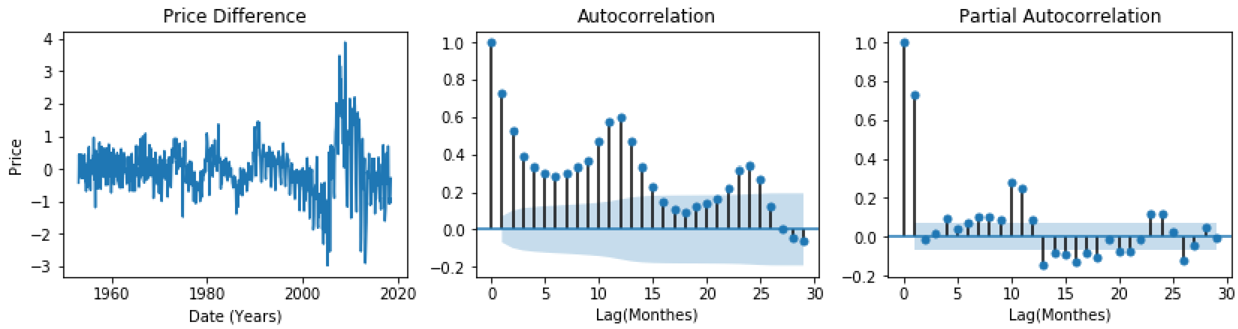
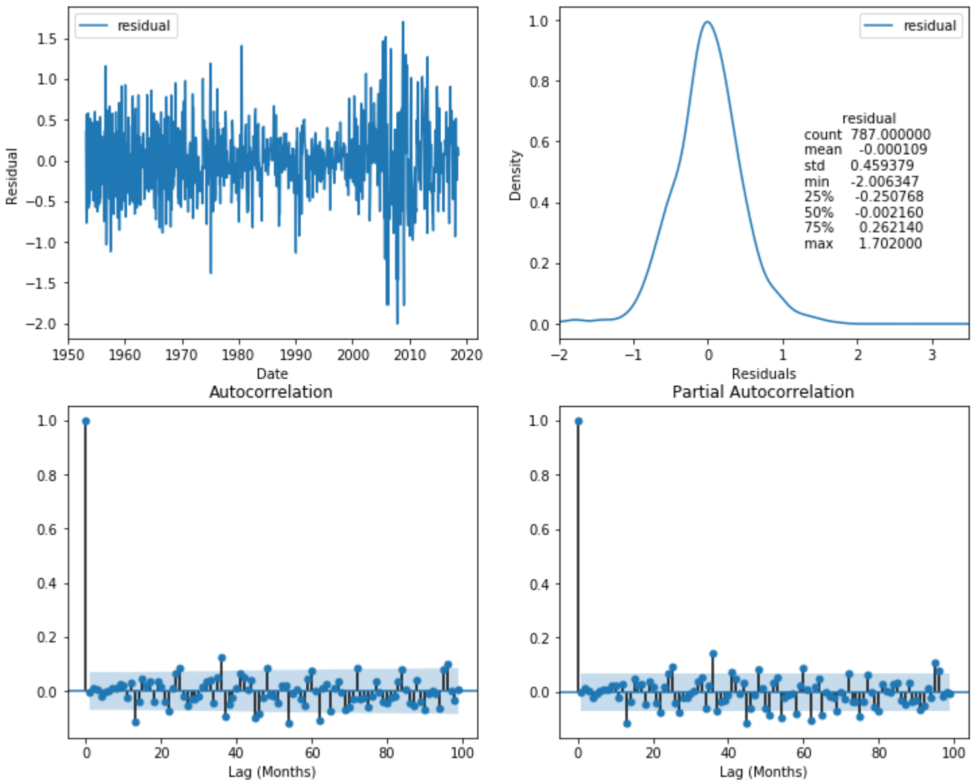


Fig. 5. Difference of Price and its Autocorrelation and Partial Autocorrelation

From the above figure, we can know that, the real home price is stationary and should be well described with an ARMA process. Thus, the integral parameter i for the ARIMA model 1. Also, from this figure, we can know the range of p,q should be both in the range from 1 to 15. By maximize the BIC score, we find the ARIMA(12,1,2) could best describe the home price. The residue of the model is shown as below:



```
        residual
count  787.000000
mean    -0.000109
std      0.459379
min     -2.006347
25%     -0.250768
50%     -0.002160
75%      0.262140
max      1.702000
```

Fig. 6. Residue of the ARIMA model and its statistical information

From the above figure, we can see that, the residue of our fitting is almost a white noise, which validate our parameter choice. The prediction of the price is shown as below:

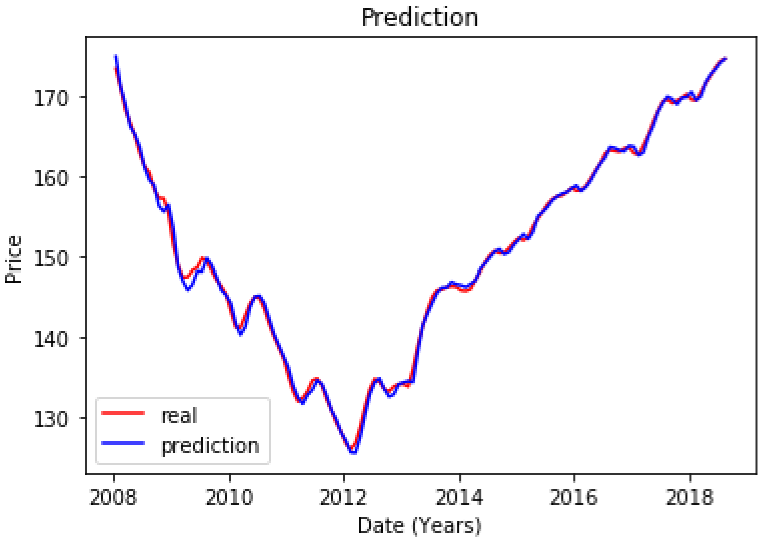


Fig. 3. Performance of ARIMA on Test Set

Interestingly, the standard deviation on the prediction of the test set is 0.86, which is better than the Xgboost result. After a careful review of the ARIMA, we found that, the AR parameter p is 12 months, exactly one year. However, in the dataset of Xgboost, there are just various macro economic factors. Seasonality is not considered in this model. Thus, this results indicate, actually, the influence of seasonality is have more significant influence on the price than other macro economic parameters. Since the seasonality is independent factor from the macro economic factors, this conclusion does not affect the importance ranking of those macro economic factors. However, in the real application, we suggest the model should include features explicitly related to seasonality.

State level:

1. Data Prepossessing and Preliminary EDA Analysis

In order to build a statewide model for real estate price dynamics, along with the above national global macro model, we need to construct a statewide data frame that incorporates all the relevant statewide Census, IRS and crime characteristics with the large Metro Case-Shiller datasets. Therefore significant effort is put into reading, parsing, transforming and merging datasets from different sources in different format, and lastly organize them by states. The datasets range from 1984 to 2017. Because we have the Case Shiller data dated back to 1970 for Atlanta-GA, Chicago-IL, Dallas-TX and San Francisco, CA, we create a different date range 1970-2018 for these 4 states. After we merge the datasets, we also preprocess some of the variables. For example, for the state total resident population that is not a ideal factor since we care more about the population density, we thus divide it by the state total area and calculate its population density relative to the total U.S population. Next, we either take the percentage or take the difference of all the variables to make them look stationary, so that they could be compared and could be used for model fitting and prediction. For our model building purpose, we restrict our scope to those states that has Metro city listed in the Case-Shiller 20 metro city list. We organize the dataset by the two-letter state abbreviation.

Next step is to visualize the dataset for each state. We want to figure out features that correlate the most the ZHVI. Before that we should have an idea of what the level data of each feature looks like. The following figure plot all the features that we feel are significant in explaining and prediction real estate price.

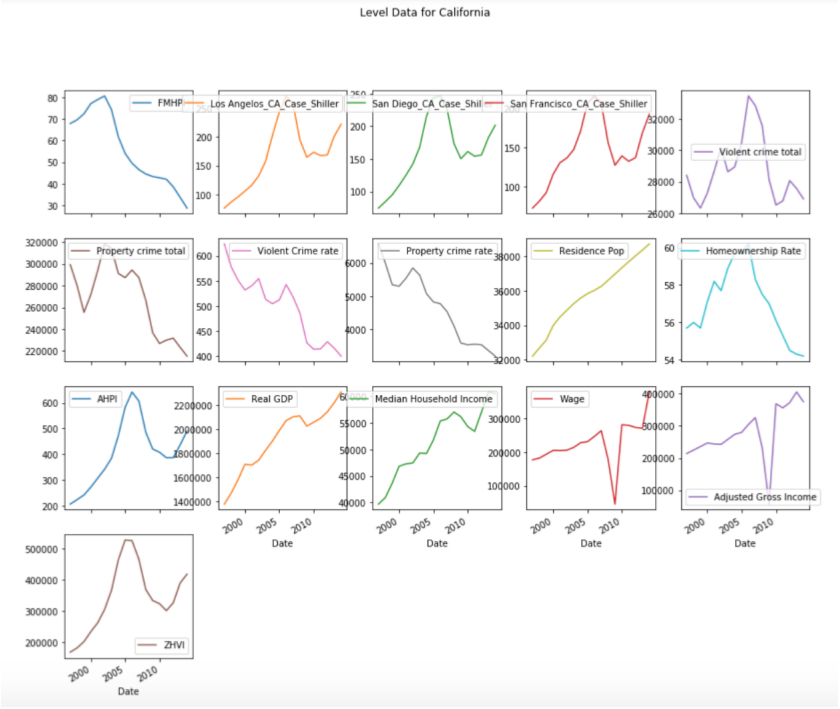


Fig: Visualization of feature level data for CA

Next, we should examine the correlation of those features with ZHVI for each state individually. The best visualization tool for this is the heatmap. From the figure below, we can see that most of the features do have a strong correlation with ZHVI and we want to pick those that has correlations that are greater than 0.5 for majority of the states as potential risk factors or alpha factors.

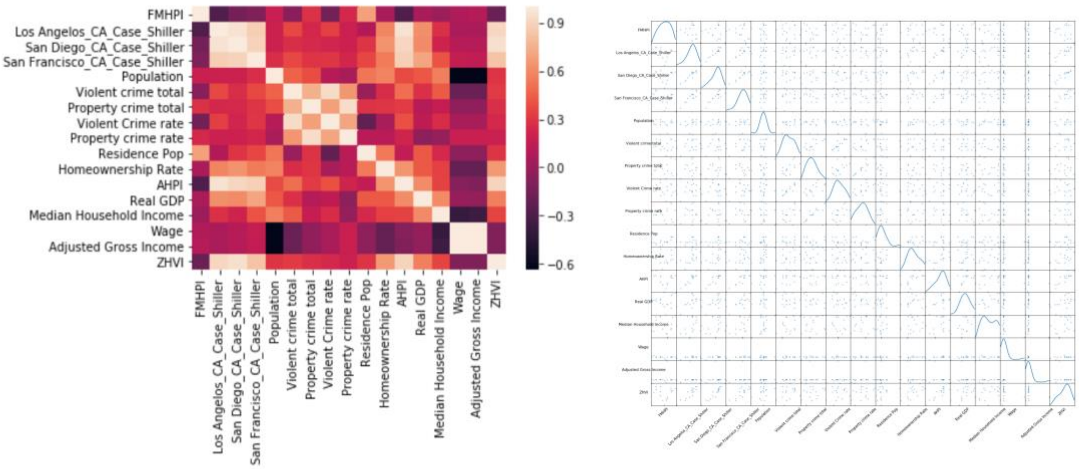


Fig: Feature Scatter Matrix and Heatmap for California

Note that in the above plot, the feature “State Minimum Wage ” has the weakest correlation with the target variable. The main reason for this is the proportion of missing values is large and we manually fill the missing values using EN-lite linear regression. Therefore we discard the feature. Also real estate price index such as AHPI and Case-Shiller tracks ZHVI closely. Since these index, especially Case-Shiller is lagged, we should not use them to build our alpha model.We would illustrate its prediction power in the ZHVI calibration section. Among the remaining features, we choose the features that have high correlation with target variables in all the states as potential risk factors. The identified significant risk factors include: Residence population density relative to U.S total population, Crime(violet) rate, Crime(property) rate, Wage & Salaries, Homeownership rate, real GDP and Median Household Income. Also, notice that some of the identified factors are regarded as "distressed" proxy factors. All the detailed implementation is in State Level Data Gathering.ipynb.

2. Modeling and Evaluation

a. OLS Benchmark Model

We build a OLS model using the risk factors identified in the previous section, plot its fitting and calculate the corresponding R-square and mean squared error.

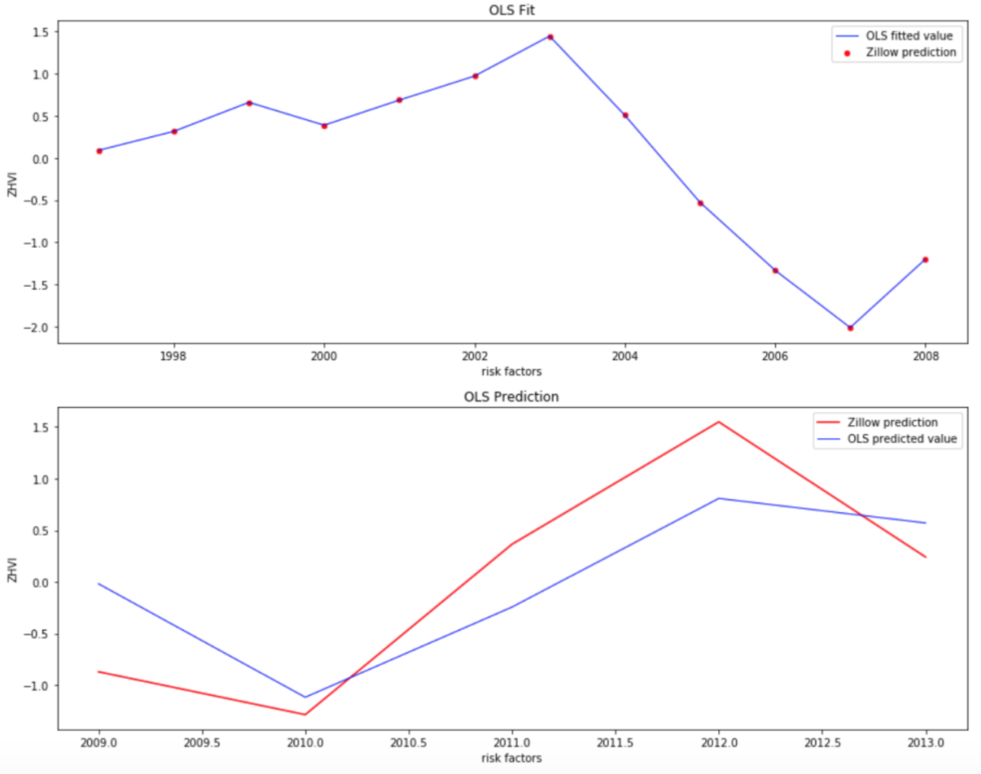


Fig: OLS fitting and prediction for California

From the plot we can see that the risk model factors explains the current co-movement of real estate price very well with approximately **64%** explained variance score. However, from the prediction, we can see that not all of the risk factors have predictive power, i.e, explains any of the alpha. We then generate the statewide prediction plot, and can come to the conclusion that not all of them are alpha factors. Then we rank the risk factors by significance score, and thus obtain the potential alpha factors. They include economic/demographic factor such as Residence Pop, Violent Crime rate and Homeownership Rate, and “distressed proxy factors” such as real GDP and Median Household Income, and we make predictions based on that. Note that all variables are quoted in return(%change or difference). The result looks better, and the following is the prediction plot for all the states.

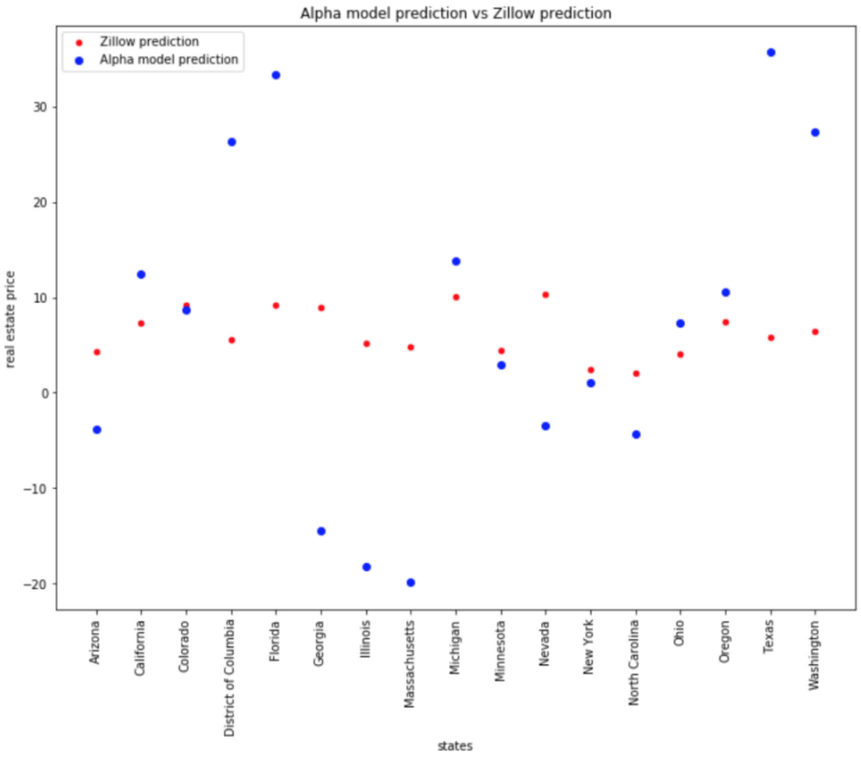


Fig: Alpha model prediction for all states

The prediction looks better than the previous model, but has place for improvement. Let's proceed to the next section.

b. OLS Model w/ Momentum factor

We want to see if momentum has risk premium, i.e, has predictive power, because investors will tend to buy house that already looks expensive. However, here we have to be cautious because house is a consumption asset. Therefore sometimes it is hard to determine whether the alpha factor is due to risk premium or due to consumption demand. Here in this case, momentum can also be explained by irrational investor biases such as herding in the presence of perceived opportunity loss, because house is a consumption asset and areas with high price momentum will typically also have high consumption demand due to better jobs, schools, household income and less crime rate.

Therefore, we check the alpha factors individually to see if they has momentum. We do this by first calculating the momentum of each alpha factor(dividing today's value by last year's value on a rolling basis), and then regress it with the target variable to check the correlation. We find out that the momentum of Median Household Income and real GDP has predictive power. In addition, we also include an indicator variable that has 1 if the previous real estate price(real estate price at time t-1) exceeds the past 3 years' average real estate price. We include these additional factors and produce the following prediction plot for all state. From the prediction result, we could see that it is better than the original model.

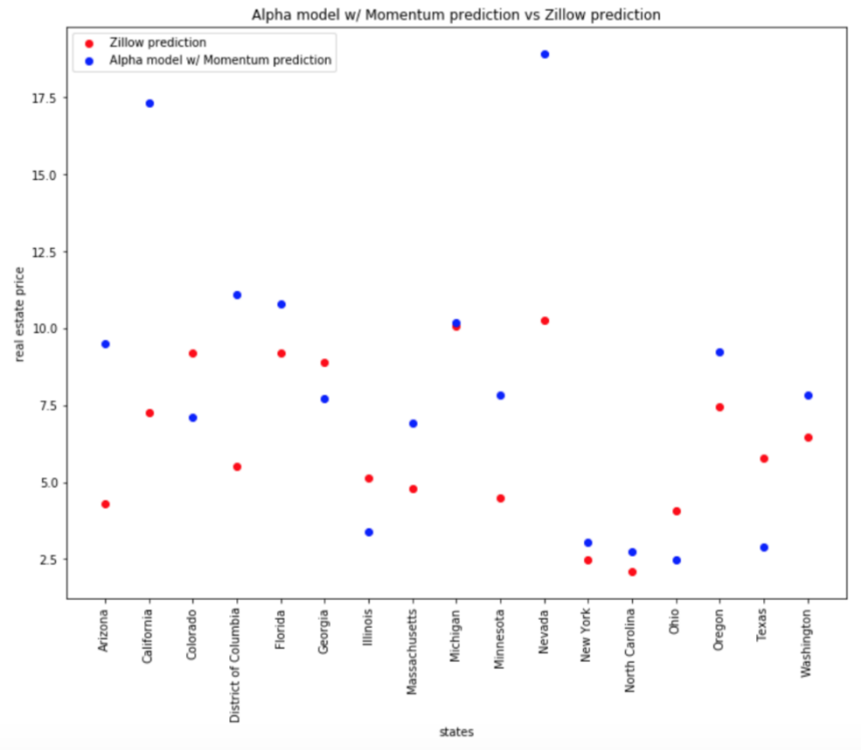


Fig: Alpha model w/ Momentum prediction for all states

c. Gradient Descent Decision Tree

We also want to train our dataset using Regression Tree model, and here we use Python’s xgboost package. Xgboost basically has Gradient Descent Decision Tree as its underlying algorithm and could be used in both classification and regression problem. In the previous OLS model, overfitting might be an issue, especially after adding the momentum factors. Therefore we hope to use xgboost and its regularization parameter to circumvent this problem. First of all, we use stratified sampling for real GDP because we notice that this feature is imbalanced. We classify the feature into 5 different range and sample proportionally within each range. Secondly, we apply Grid Search Cross Validation method to find the optimal parameters by looking at it learning curve. If the learning rate is smaller than 0.1, we would like to add more weak learners. After we find the optimal parameters, we can proceed to fit the training data and get the prediction. Below is the output the xgboost method.

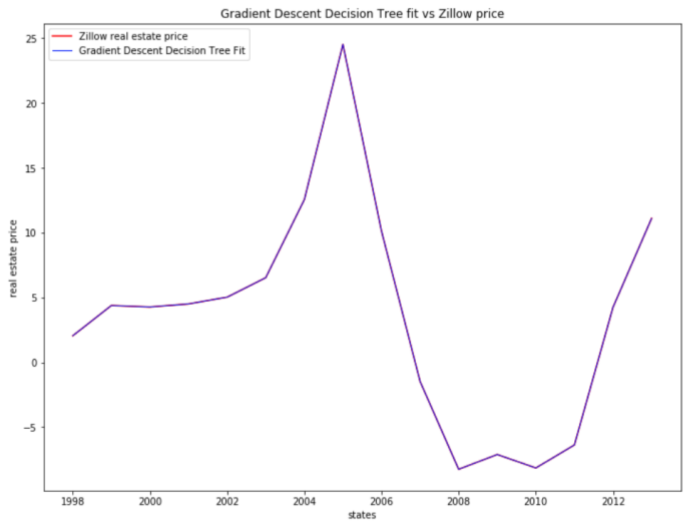


Fig: Xgboost fitting for California

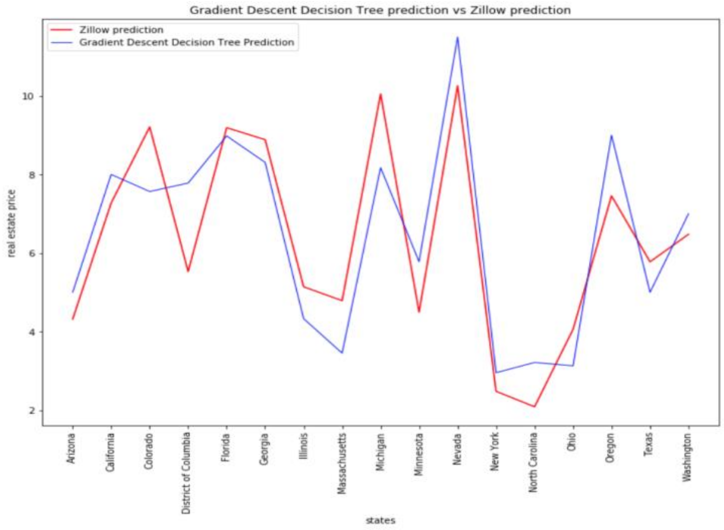
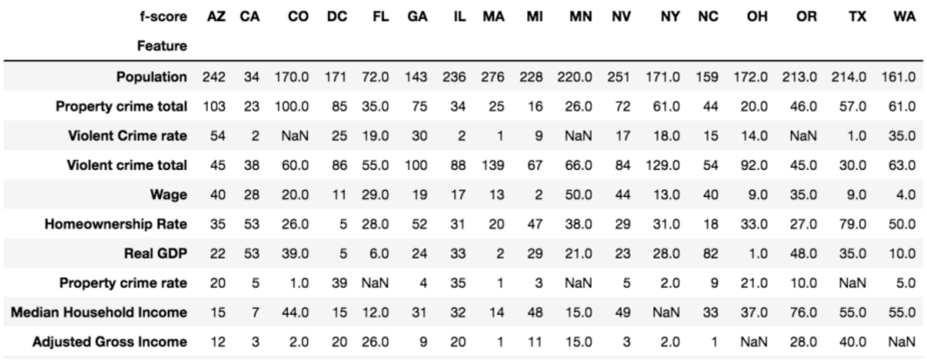


Fig: Xgboost prediction for all states

We could see that the prediction is fairly good. Therefore we successfully build a statewide covariance model and an alpha model with an explained variance score of **77.31%**. We close our discussion of statewide model with a table of factor score. All the detailed implementation is in State Level Feature Engineering.ipynb.



f-score	AZ	CA	CO	DC	FL	GA	IL	MA	MI	MN	NV	NY	NC	OH	OR	TX	WA
Feature																	
Population	242	34	170.0	171	72.0	143	236	276	228	220.0	251	171.0	159	172.0	213.0	214.0	161.0
Property crime total	103	23	100.0	85	35.0	75	34	25	16	26.0	72	61.0	44	20.0	46.0	57.0	61.0
Violent Crime rate	54	2	NaN	25	19.0	30	2	1	9	NaN	17	18.0	15	14.0	NaN	1.0	35.0
Violent crime total	45	38	60.0	86	55.0	100	88	139	67	66.0	84	129.0	54	92.0	45.0	30.0	63.0
Wage	40	28	20.0	11	29.0	19	17	13	2	50.0	44	13.0	40	9.0	35.0	9.0	4.0
Homeownership Rate	35	53	26.0	5	28.0	52	31	20	47	38.0	29	31.0	18	33.0	27.0	79.0	50.0
Real GDP	22	53	39.0	5	6.0	24	33	2	29	21.0	23	28.0	82	1.0	48.0	35.0	10.0
Property crime rate	20	5	1.0	39	NaN	4	35	1	3	NaN	5	2.0	9	21.0	10.0	NaN	5.0
Median Household Income	15	7	44.0	15	12.0	31	32	14	48	15.0	49	NaN	33	37.0	76.0	55.0	55.0
Adjusted Gross Income	12	3	2.0	20	26.0	9	20	1	11	15.0	3	2.0	1	NaN	28.0	40.0	NaN

Fig: f-score for xgboost for all states

County level:

Exploratory data analysis:

1. Results:

- Crime rate could predict counties with low GDP per capita. It's not a good factor for counties with high GDP per capita.
- Education: Education could not significantly account for house values
- Income: Income could significantly account for house values

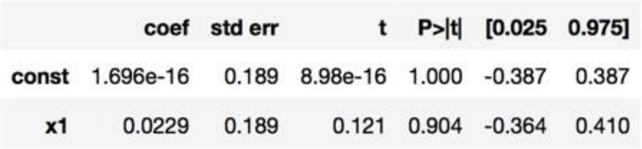
2. Methodology:

The main problem with the county prediction is that it's hard to gather time series data for all counties at all time. Thus some exploratory data analysis is necessary for us to get some insights for further data analysis.

For all our analysis, we use Zillow index ZHVI as representative for house value.

<u>2.1 Crime Rate:</u>

First, we are interested in how house value relate to crime in this area. But we don't have crime rate for all counties. So we choose some counties from high GDP per capita areas and some counties from low GDP per capita areas, and run cross-sectional OLS regression on them. The results show that for high GDP per capita areas, house value is not significantly related to crime rate. But for low GDP per capita areas, house value is significantly negatively related to crime rate. Possible reasons for that is: those high GDP places are mostly highly populated, busy and when people choose their living areas, they may consider more on traffic, children's education, convenience to work rather than crime rate.

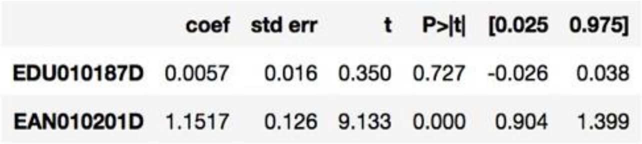


	coef	std err	t	P>|t|	[0.025	0.975]
const	1.696e-16	0.189	8.98e-16	1.000	-0.387	0.387
x1	0.0229	0.189	0.121	0.904	-0.364	0.410

Here’s a summary of regression on counties in high GDP state: CA

2.2 Education and Income

We are also interested in the relationship between education and income. We regress ZHVI of 1000 counties against education score and income.(See data description). Here’s the regression results:



	coef	std err	t	P>|t|	[0.025	0.975]
EDU010187D	0.0057	0.016	0.350	0.727	-0.026	0.038
EAN010201D	1.1517	0.126	9.133	0.000	0.904	1.399

EDU010187D: education score

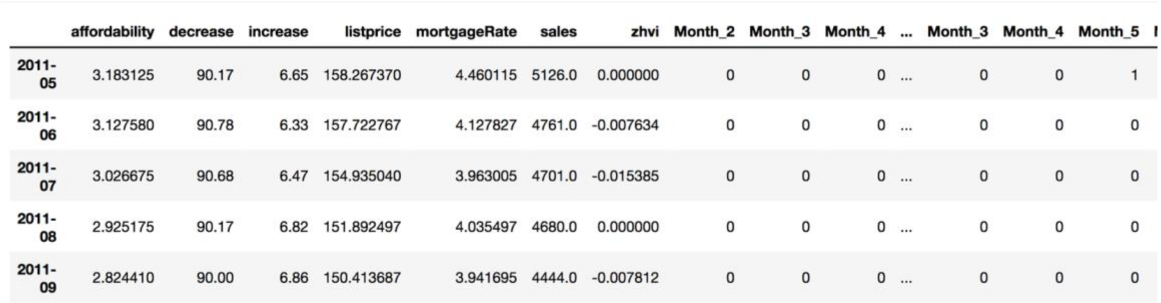
EAN010201D: income

We could see that education is not significant and income is positively significant.

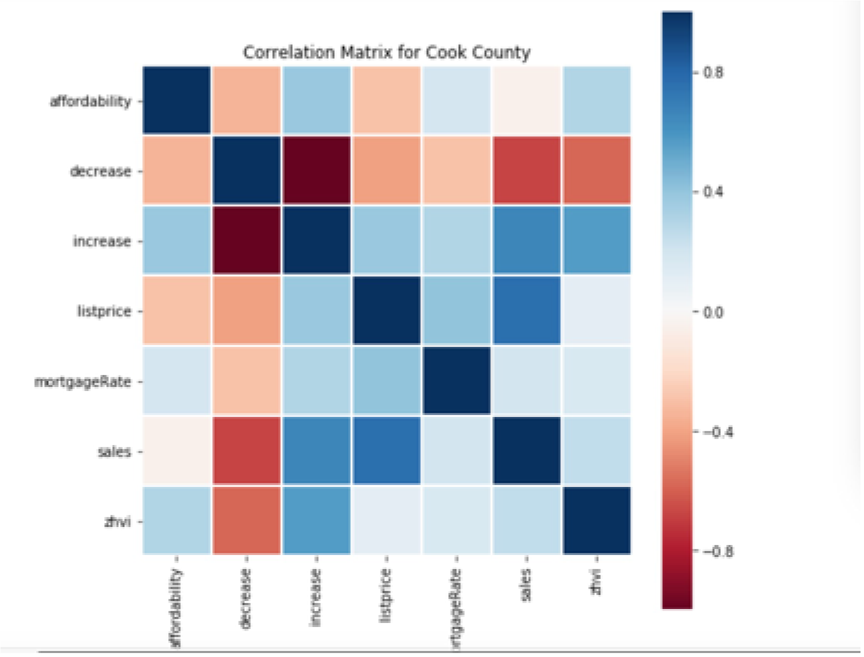
1. Factor Pool:

Our factor pool consists these factors: (All are monthly data)
We use monthly data from 2011-3 to 2018-10, since list price and sales are only available from 2011-3
Affordability: Zillow’s index for mortgage affordability. Similar to price-to-income ratio. See calculation: https://www.zillow.com/research/data/
Since affordability is state-level data, not county level. We map the affordability to each county based on which state it’s located. All counties within same state will have the same affordability data
Listprice: The percentage of current for-sale listings on Zillow with a price cut during the month.
Sales： The number of homes sold during a given month. See calculation: https://www.zillow.com/research/home-sales-methodology-7733/
Decrease: The percentage of homes in an given region with values that have decreased in the past year.
Increase: The percentage of homes in an given region with values that have increased in the past year.
Seasonal dummy variables: House prices are very time series that have seasonal trend. We use 11 dummy variables for 12 months. (Not 12 variables to avoid dummy trap)
Mortgage rate: Weekly quoted mortgage rate. We calculate the mean for each month as mortgage rate factor.

Here’s the factor matrix for Cook County and their correlation matrix:



	affordability	decrease	increase	listprice	mortgageRate	sales	zhvi	Month_2	Month_3	Month_4	...	Month_3	Month_4	Month_5	I
2011-05	3.183125	90.17	6.65	158.267370	4.460115	5126.0	0.000000	0	0	0	...	0	0	1	
2011-06	3.127580	90.78	6.33	157.722767	4.127827	4761.0	-0.007634	0	0	0	...	0	0	0	
2011-07	3.026675	90.68	6.47	154.935040	3.963005	4701.0	-0.015385	0	0	0	...	0	0	0	
2011-08	2.925175	90.17	6.82	151.892497	4.035497	4680.0	0.000000	0	0	0	...	0	0	0	
2011-09	2.824410	90.00	6.86	150.413687	3.941695	4444.0	-0.007812	0	0	0	...	0	0	0	



2. Benchmark: OLS

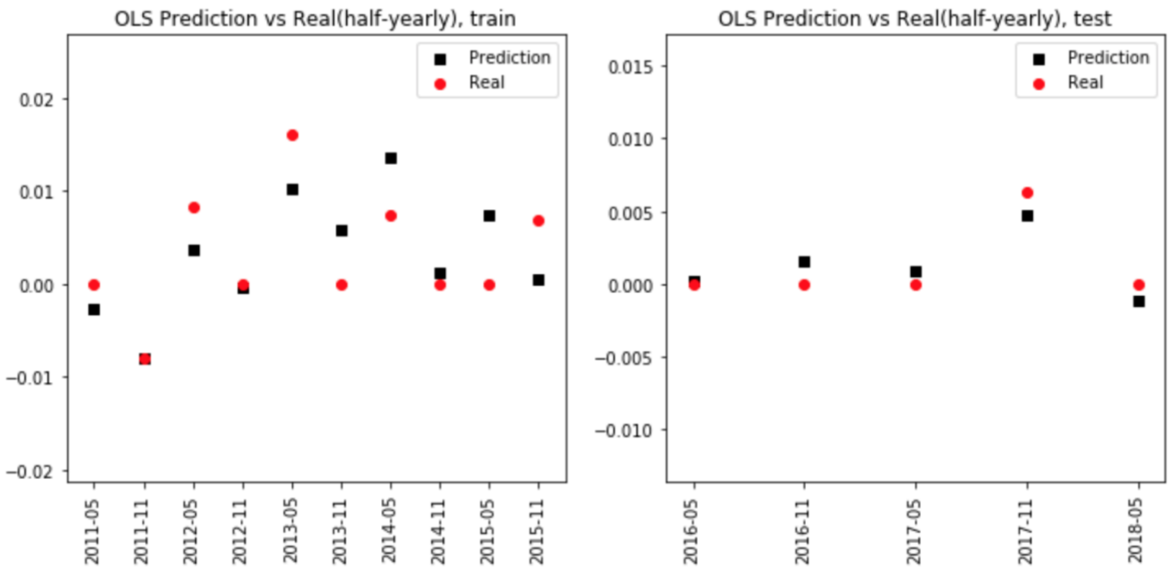
2.1 Cook county prediction

We use cook county in Chicago as an example. We use the first 2/3 of the data as training set, and the rest 1/3 as testing set.
Here’s the prediction results for cook county: (We present half-yearly data)

Where:

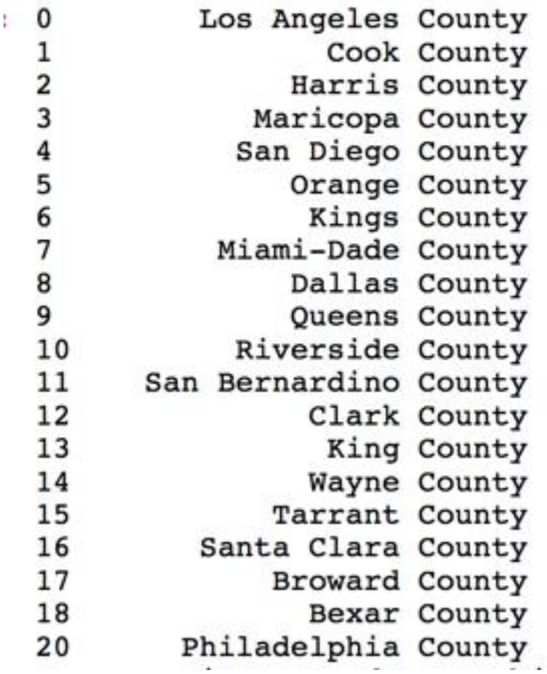
corr_train(prediction, real) = 0.6966946

corr_test(prediction, real) = 0.88670395



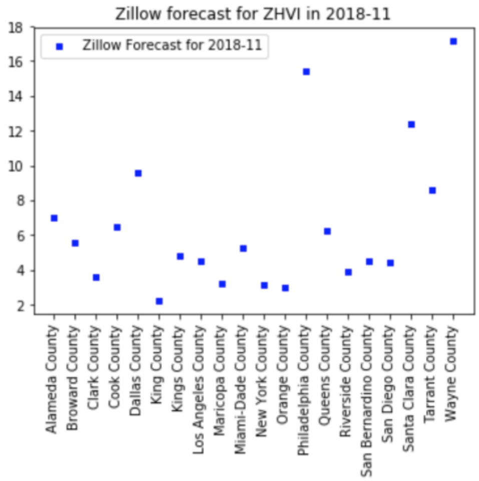
2.2 Prediction for the largest 20 counties and comparison with Zillow forecast

In case of missing data, we use the largest size 20 counties for our prediction. We choose the top size rank counties in zillow's ZHVI dataset. Here are the 20 counties:

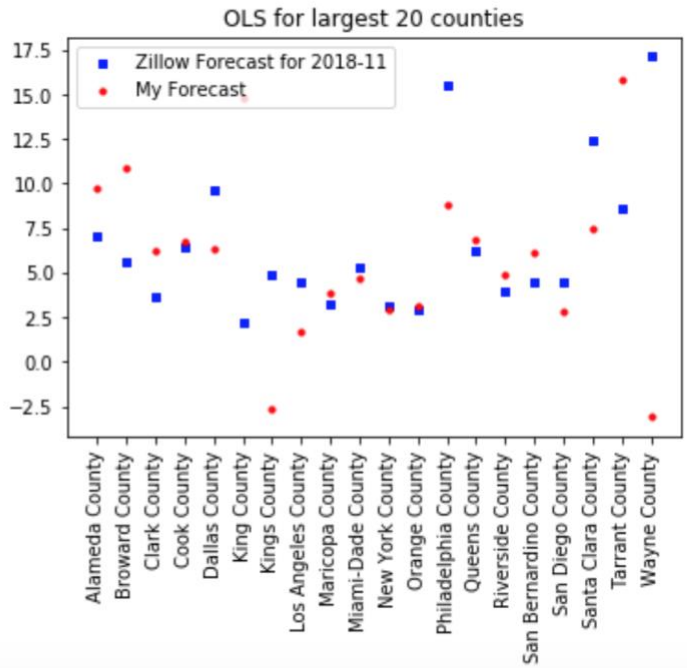


```
0         Los Angeles County
1                Cook County
2              Harris County
3            Maricopa County
4           San Diego County
5              Orange County
6               Kings County
7          Miami-Dade County
8              Dallas County
9              Queens County
10          Riverside County
11     San Bernardino County
12              Clark County
13               King County
14              Wayne County
15            Tarrant County
16        Santa Clara County
17            Broward County
18              Bexar County
20       Philadelphia County
```

Zillow forecast is Zillow research's forecast for the next period ZHVI, as percentage change.



We only have Zillow forcast of ZHVI for 2018-11 as percentage change at the time we start our project. Thus we compare our training results with Zillow forecast for 2018-11. Training horizon: 2011-5 to 2018-10:

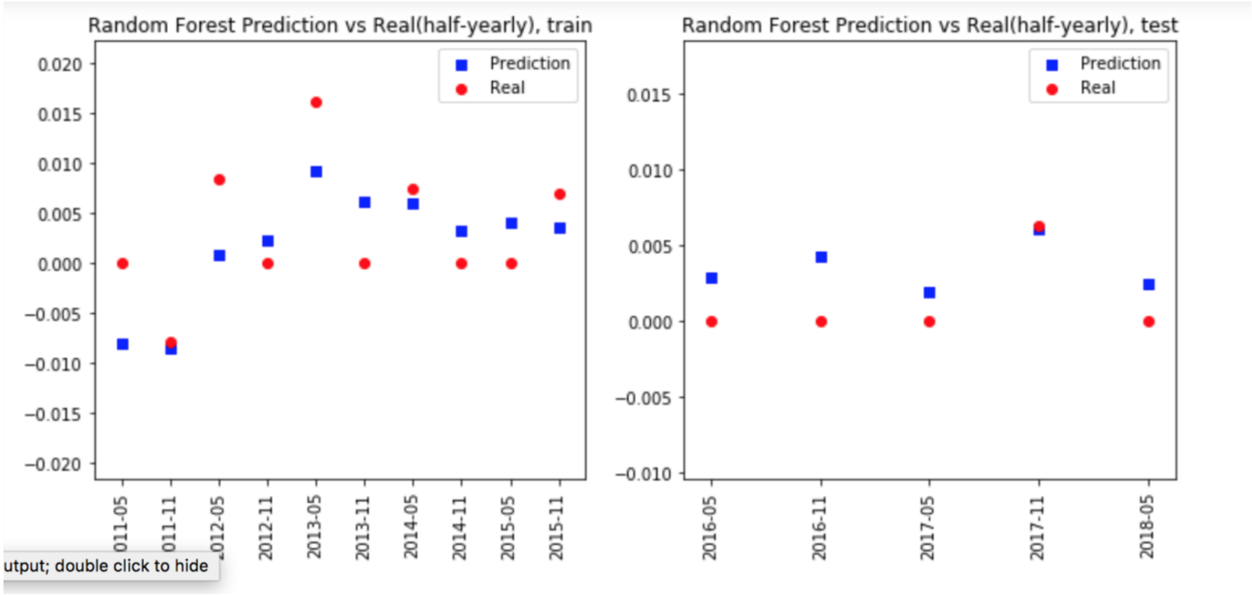


Goodness of prediction: For the 20 counties, 7 counties have the difference > 5.5%.

3.Prediction: Random forest

We are also very interested in how good random forest could predict zhvi.
We use the same training set and testing set.

corr_train(prediction, real) = 0.67062471 corr_test(prediction,real) = 0.85602227



Some comments: Due to lack of data for counties, our factor pool only consists 6 factors, and the prediction power is not strong. Therefore we could make more efforts on getting enough data.

6. Conclusions

For the real estate price project, we successfully accomplish (1) the calibration of ZHVI to Case-Shiller and investigate the potential relations between the two indices. (2) build and train the national global/statewide/countywide real estate price model, identify risk and alpha factors and obtain prediction respectively. One of the main challenge leftover would be the fact that real estate is a consumption asset, and thus after we find out those factors, it is still hard to distinguish if it is due to risk premium or due to consumption demand.

In addition, there is a lot of space for improvement for our training data set. For example, we could include data on immigration patterns. We could also make efforts to scrap more data to increase the length of our training set. For now, the different features mostly have inconsistent time frequency and we either use extrapolation to fill the missing value or cut the data short, which is also a source for future improvement.

7. References

[1] Case-Shiller index methodology.
https://us.spindices.com/documents/methodologies/methodology-sp-cs-home-price-indices.pdf

[2] ZHVI index methodology. https://www.zillow.com/research/zhvi-methodology-6032/

[3] PRICES OF SINGLE FAMILY HOMES SINCE 1970: NEW INDEXES FOR FOUR CITIES. https://www.nber.org/papers/w2393.pdf