Real Estate Price Modelling

Siyue Su

Sulin Liu

Bixing Yan

Motivation & Background

- Main Objective: Discover potential risk factors and alpha factors
- Background: Zillow already has published a basic 12-month home value forecasting online at the State, County, Zip code and Metro level. **But** Zillow's forecasting model does not fully justify their choice of features.
- Motivation: 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.

Methodology

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

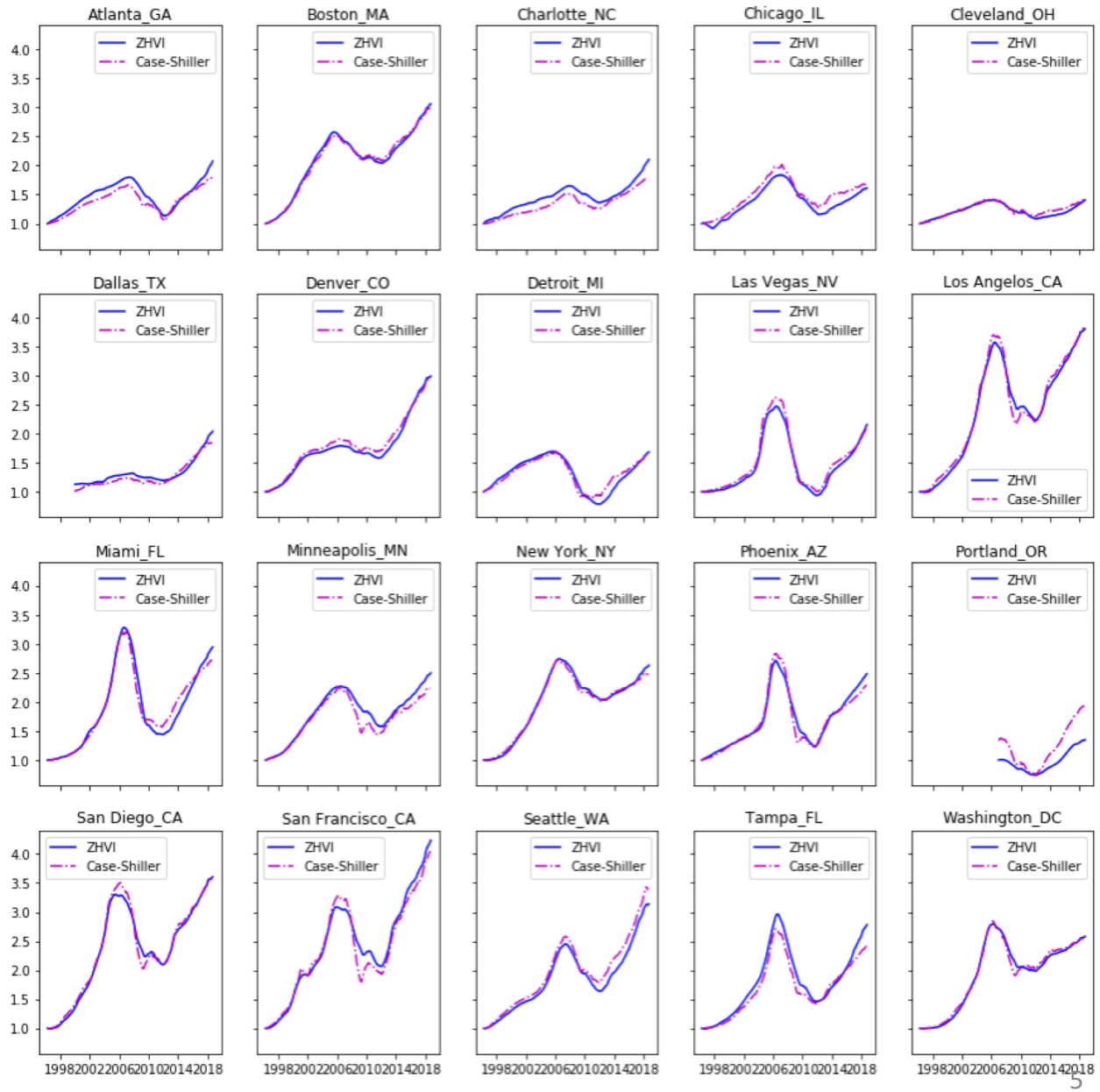
We follow this procedure for nation/state/county level models respectively

Before we go to modelling part….

- We want to first calibrate two indices that measure the real state price
- One is ZHVI: Zillow's own forecast
- One is Case-Shiller index

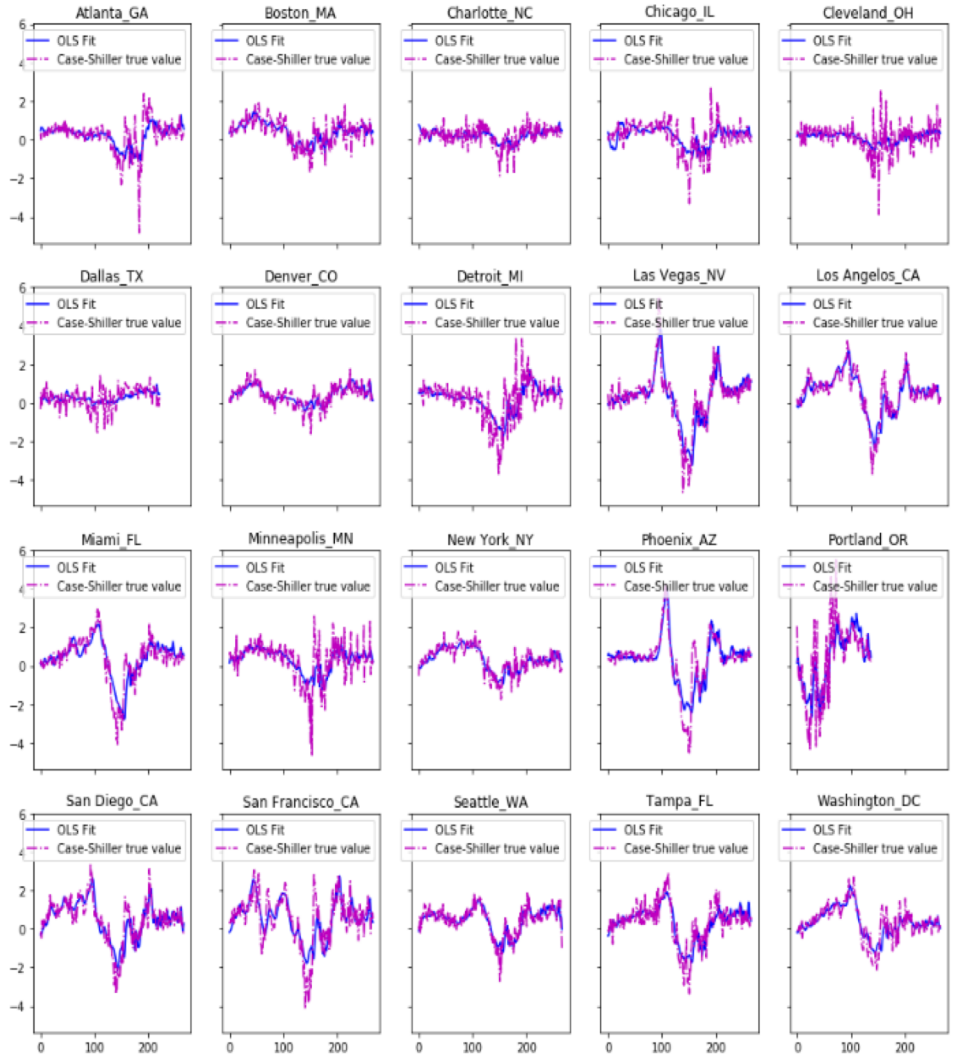
Results of ZHVI Calibration to Case-Shiller

- The result is for all 20 metro cities
- Examine and check if the two indices track and follows each other
- Is ZHVI upward-biased due to %foreclosures?



Prediction of Case-Shiller using ZHVI

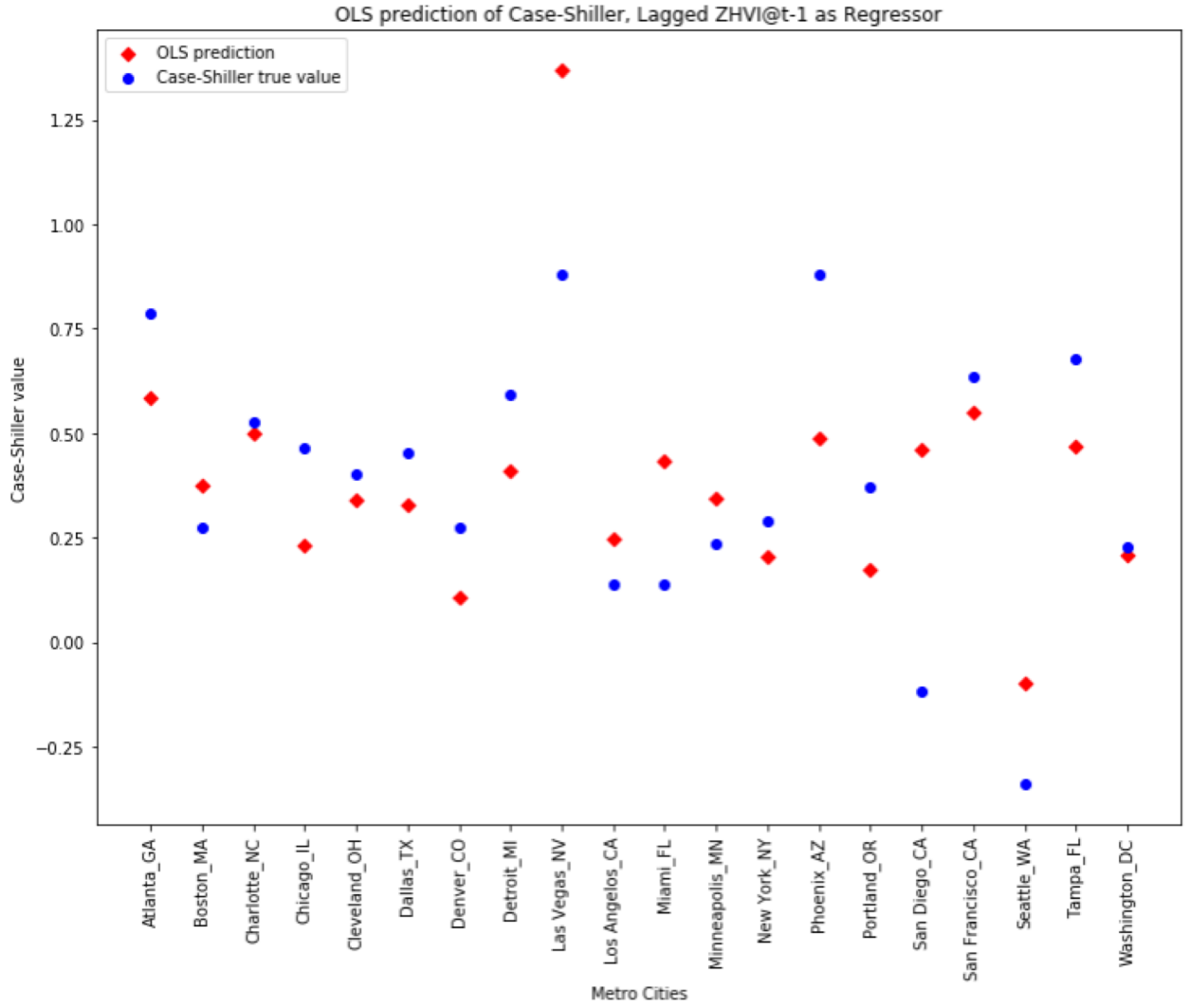
By examining the β of OLS model, there is strong indicative power of lagged ZHVI to Case-Shiller for majority of metro cities



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

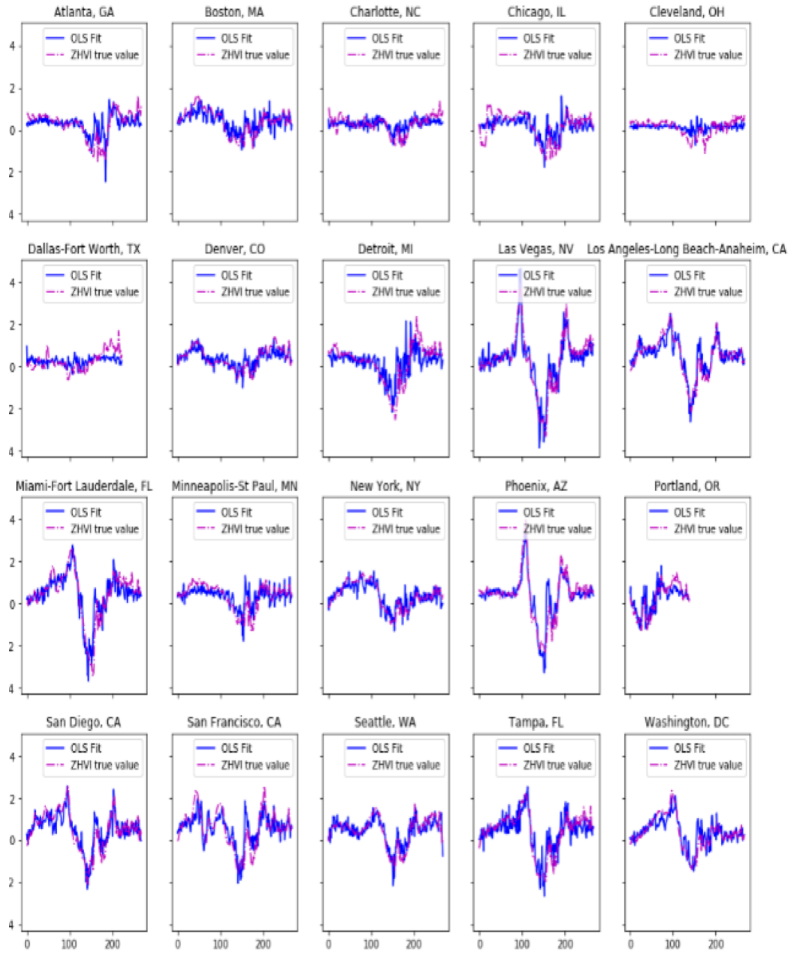
Prediction of Case-Shiller using ZHVI

The result of prediction is very good with accuracy score displayed in the previous slide



Prediction of ZHVI using Case-Shiller

- By examining the β of OLS model, there is only weak indicative power of lagged Case-Shiller to ZHVI for majority of metro cities
- The accuracy of prediction is low



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

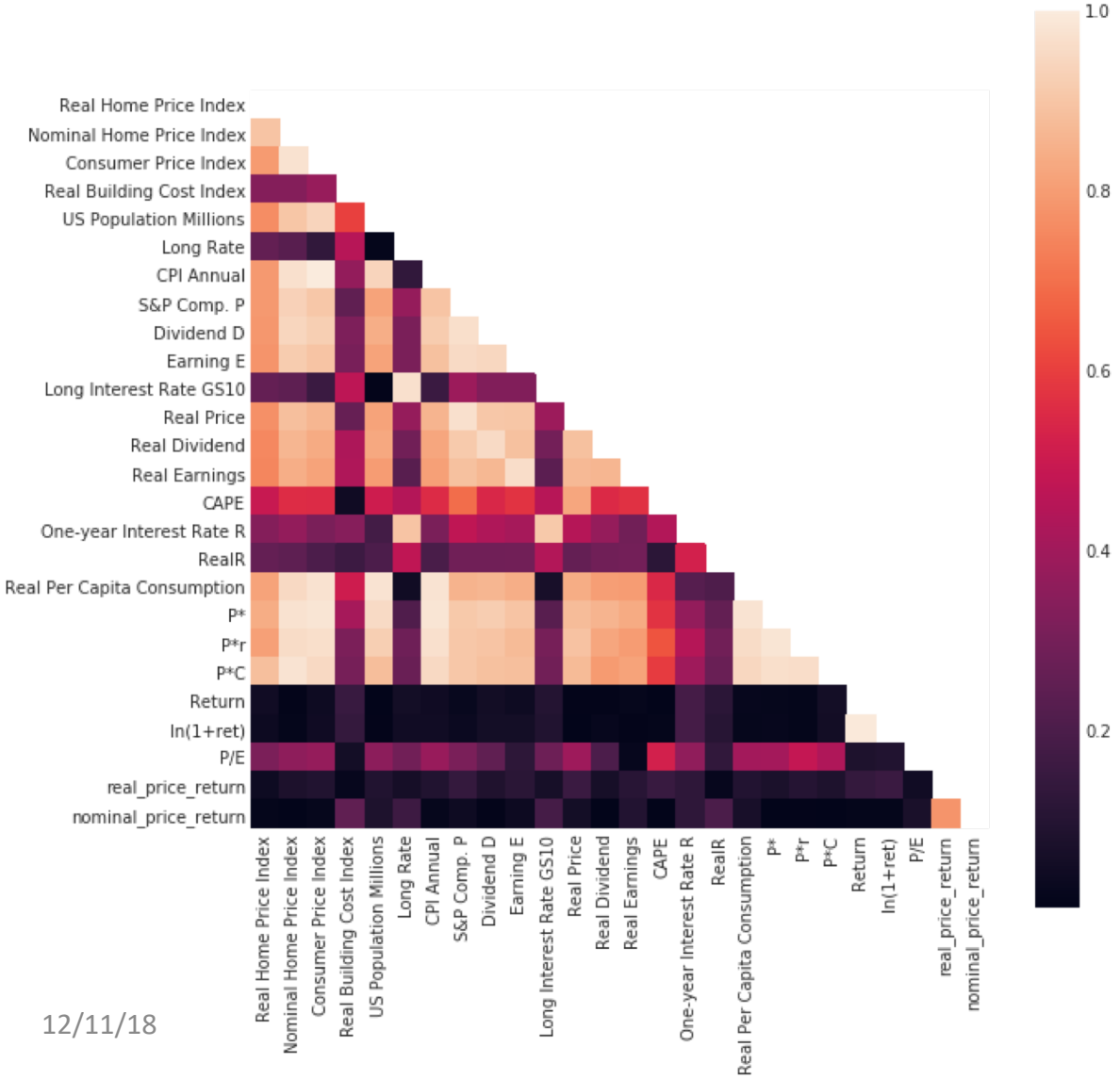
A Study of National Level House Price with Boost Tree and ARIMA

Feature Importance in National Level

National Level House Price Index

- **Real Estate Features:**
  - Building Cost Index
  - Population Level
- **Stock Market Features**
  - Stock Price Index
  - Consumer Price Index
  - Interest Rate

Which Target is Better? - Correlation



Real vs **Nominal**

- Higher Correlation
- Reflect the Nature

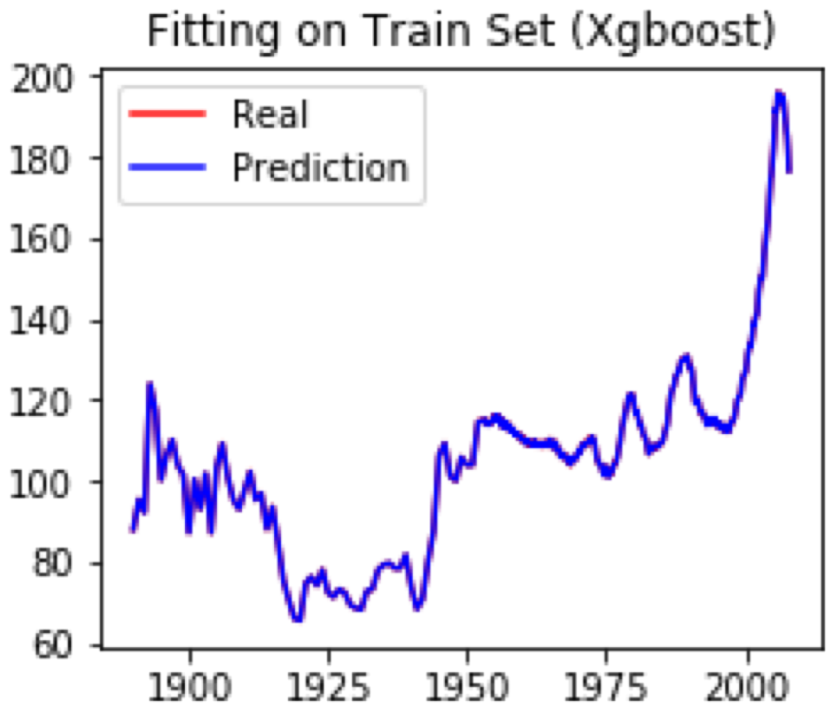
Return vs **Price**

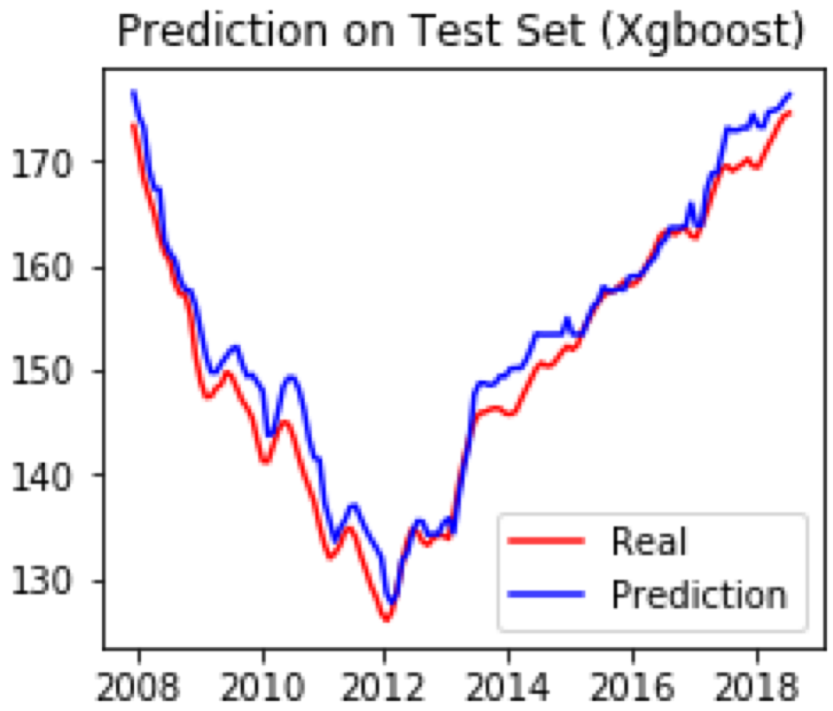
- Less Correlation
- Related to Difference?

Interpretation on Features – Tree Model

Tree Model

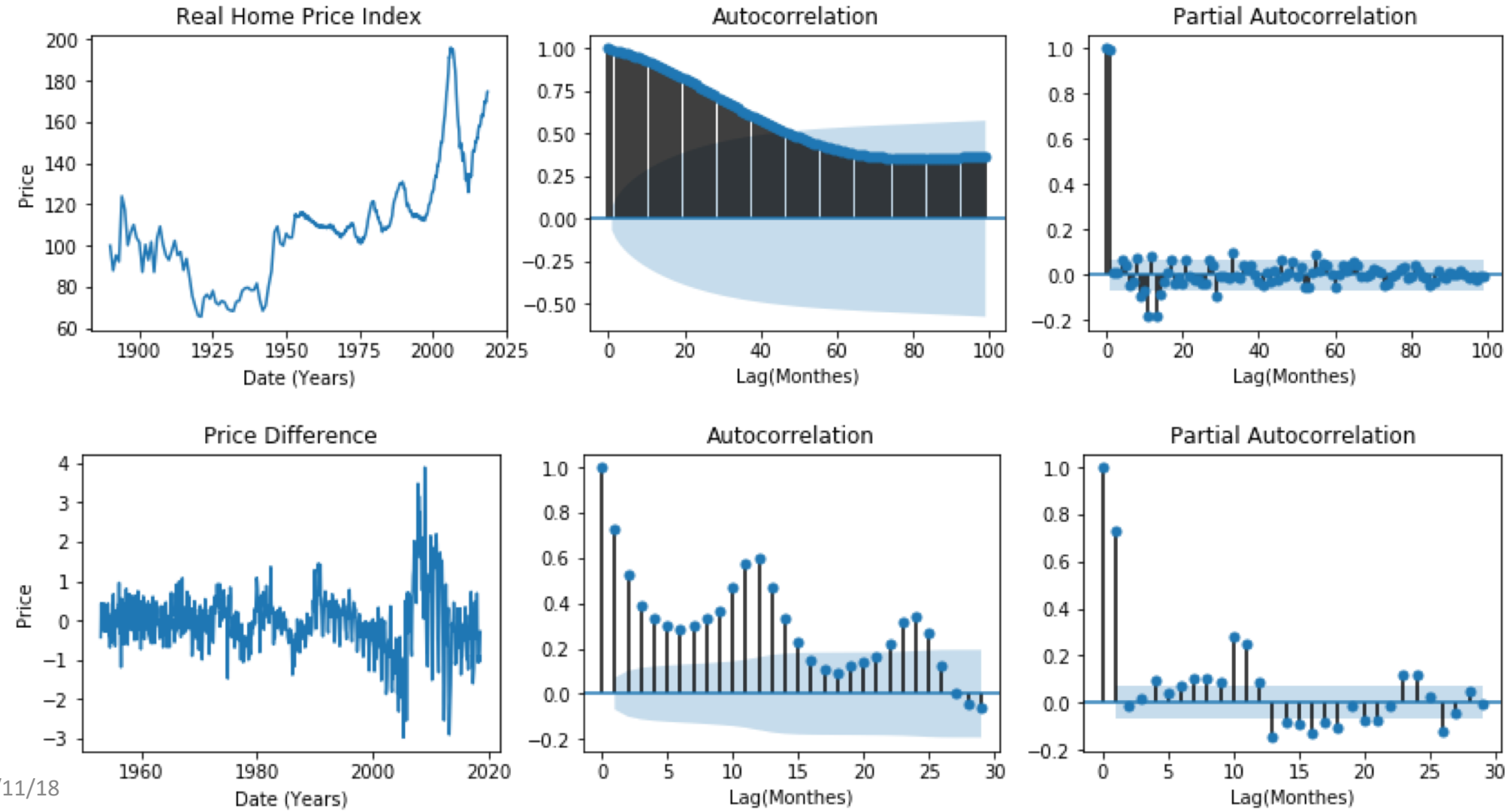
- Interpretability
- Convenience
- Accuracy
- Overfitting



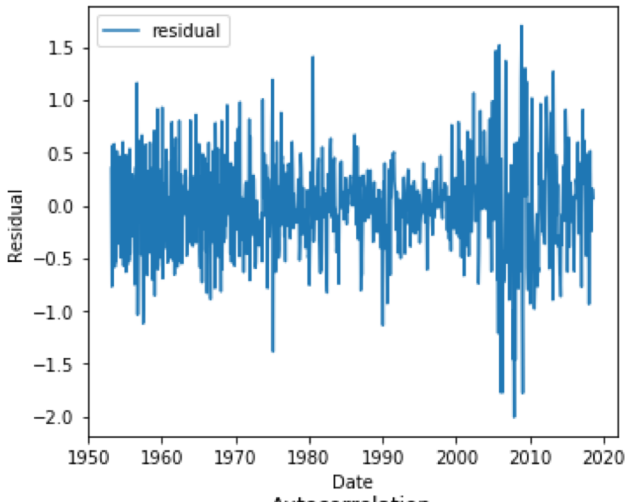


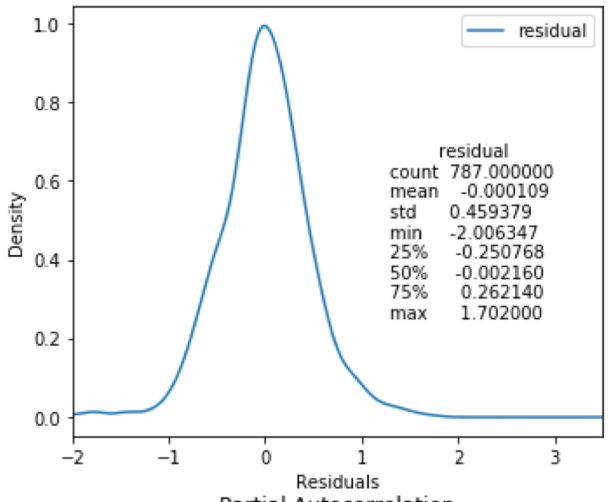
Feature	Long Rate GS10	S&P Comp. P	CAPE	CPI	Real D	Real P	Real E	E	Real Building Cost	Real R	D	R	Long Rate	P/E	P*C	P*r	Population	P*	Real PCC
f score	443	348	344	321	311	309	308	242	156	145	126	94	84	78	72	58	52	20	13

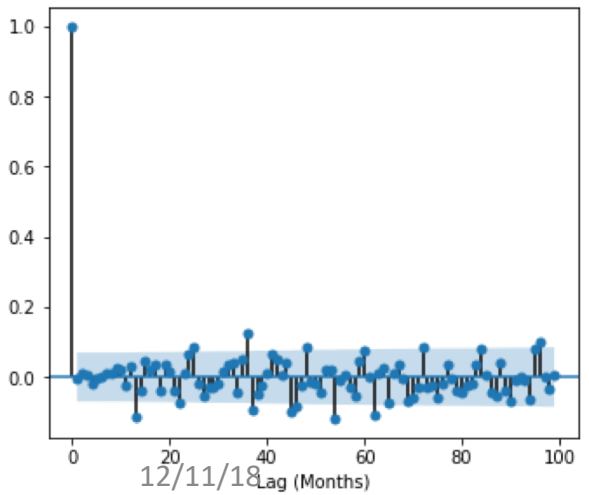
How About Traditional Method? (ARIMA)

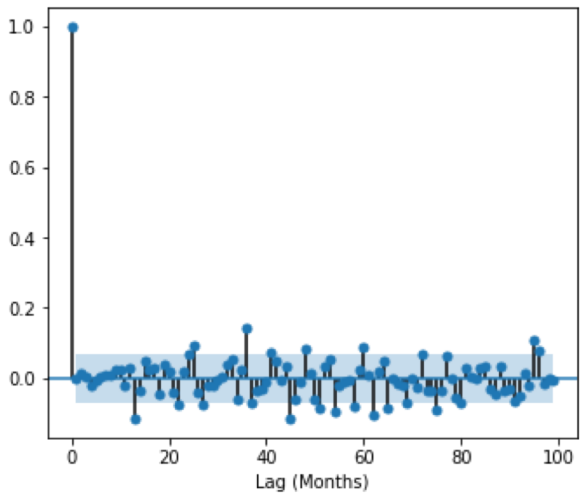


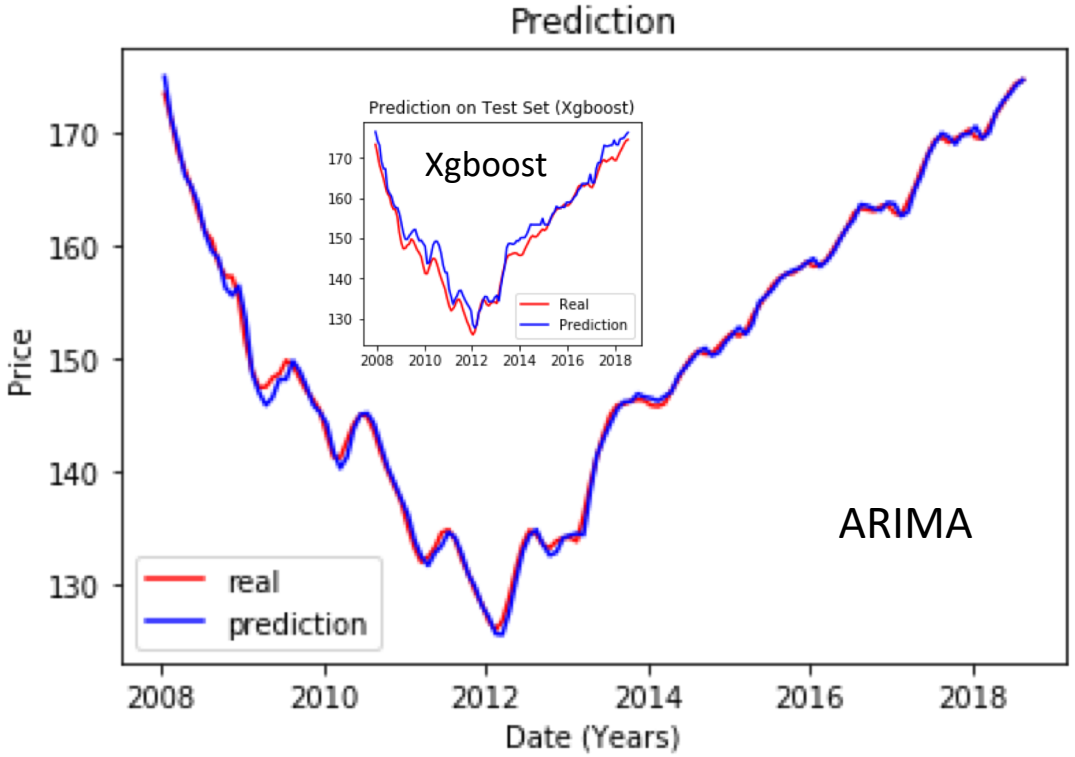
ARIMA(12,1,2) – Why A Better Model?









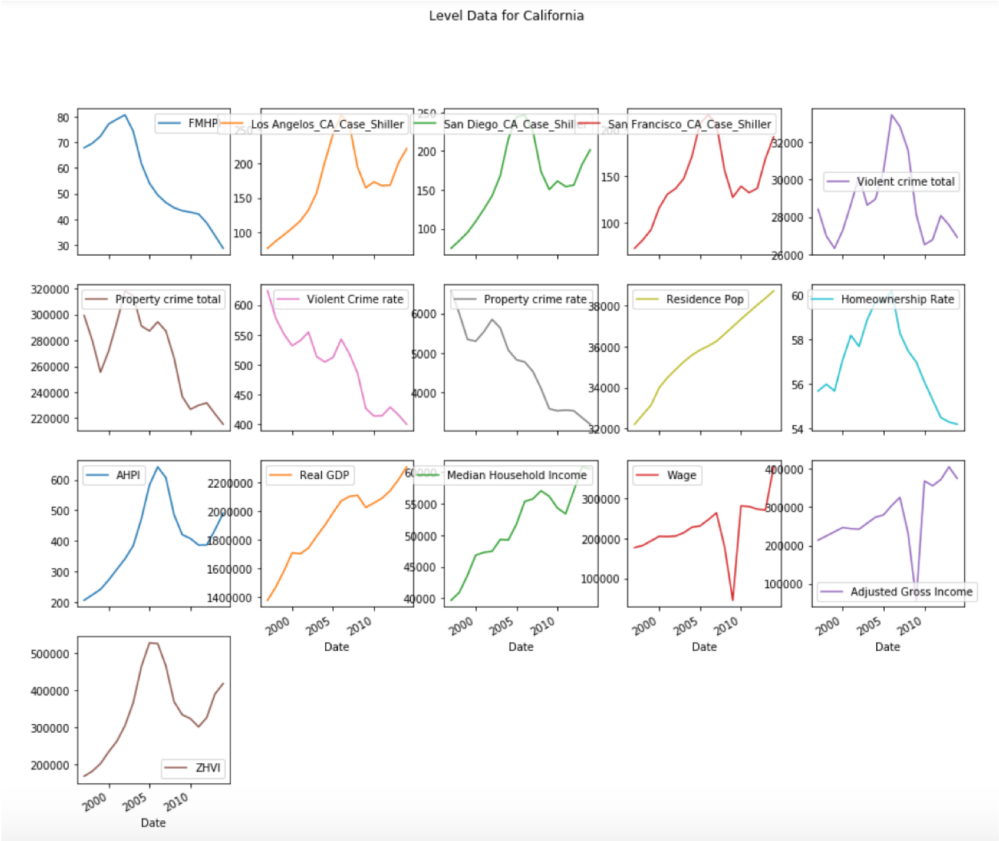


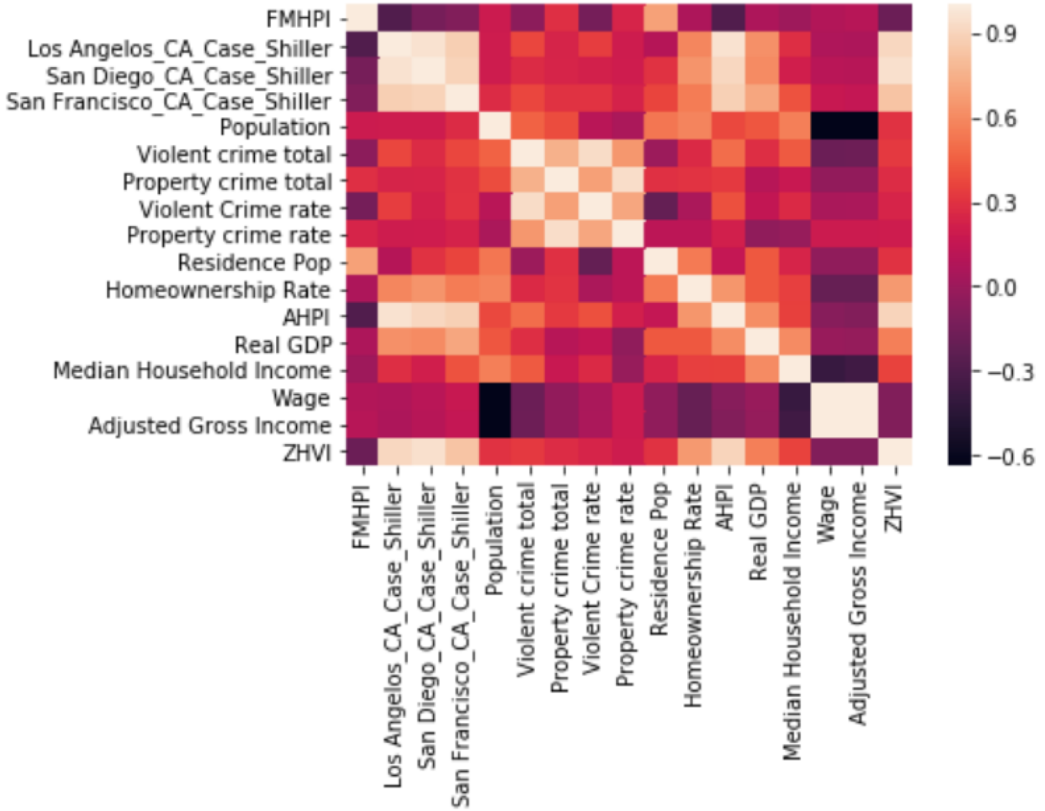
$$\sigma_{xgboost} = 2.8 \ vs \ \sigma_{ARIMA} = 0.9$$

Seasonality is more important!

A Study of National Level House Price with Boost Tree and OLS

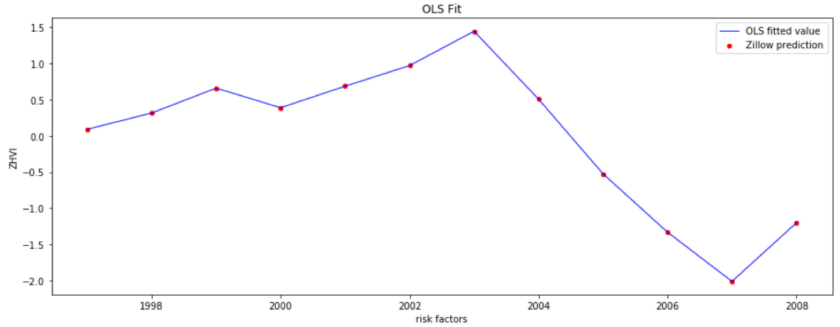
Feature Importance in State Level

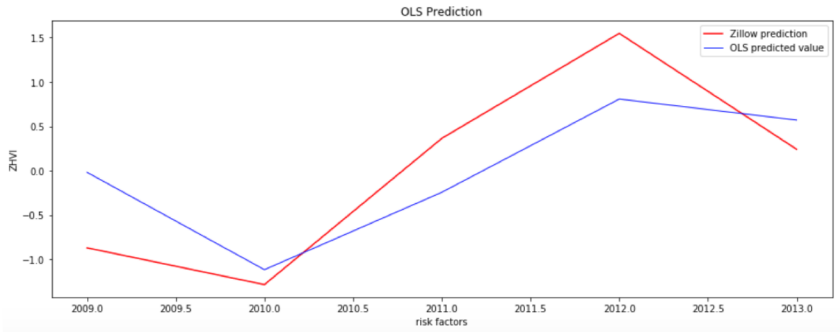


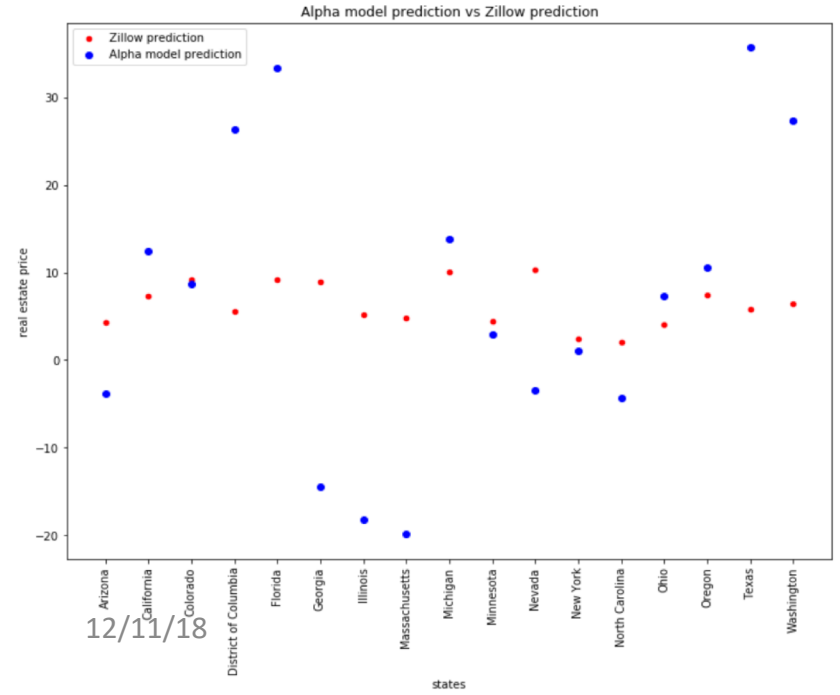


State Level Model
- Data Prepossessing and Preliminary EDA Analysis

Examine the heatmap and pick the factors with correlation higher than 0.5 as potential risk factors





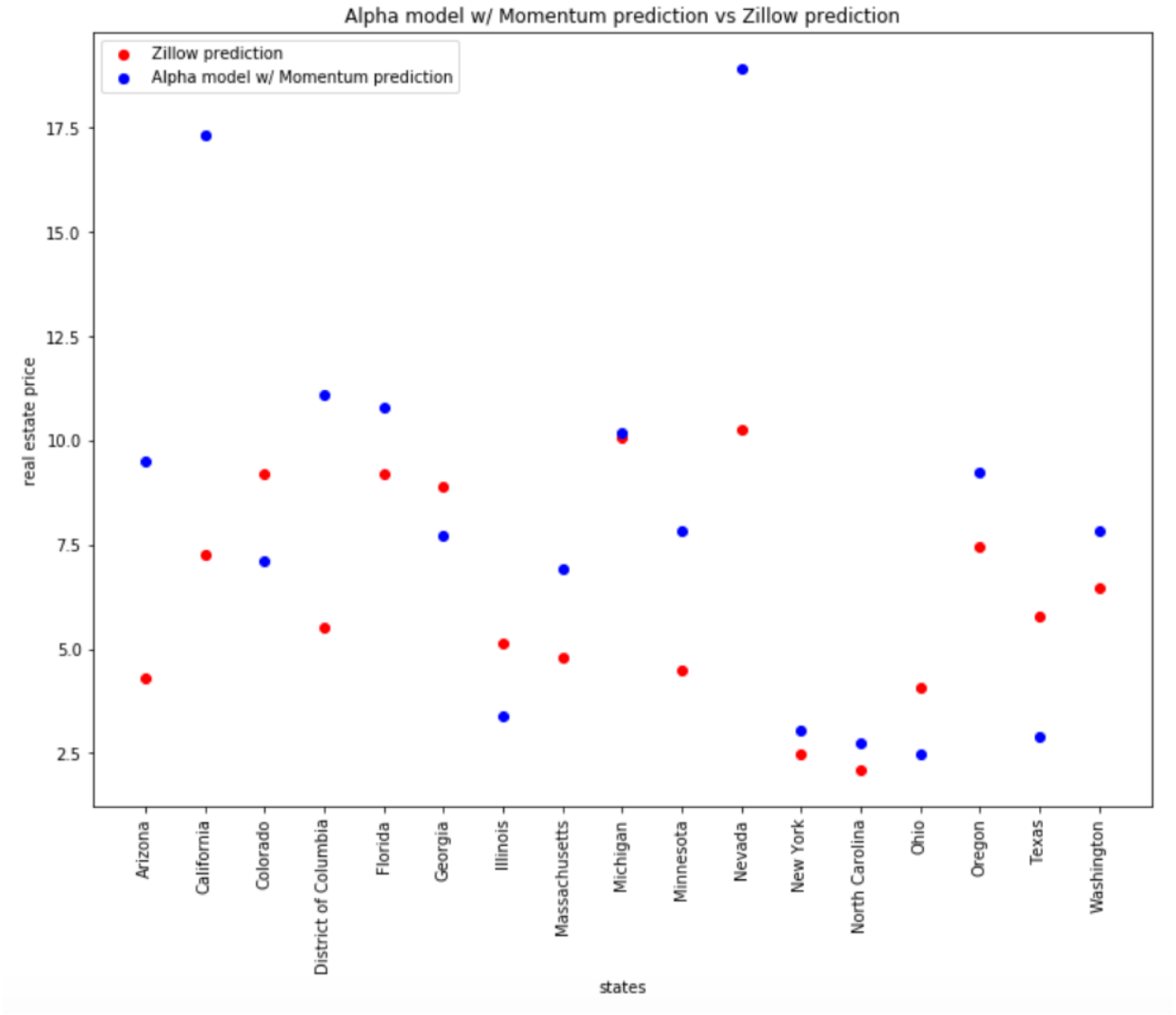


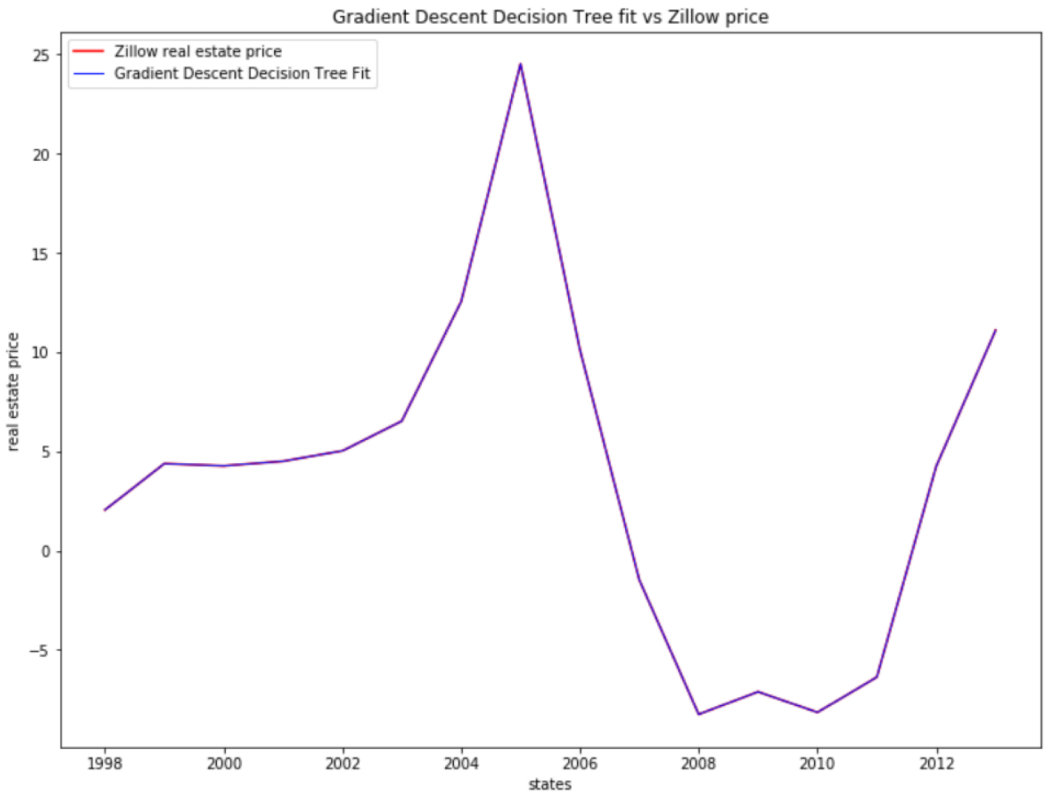
State Level Model - OLS Benchmark Model

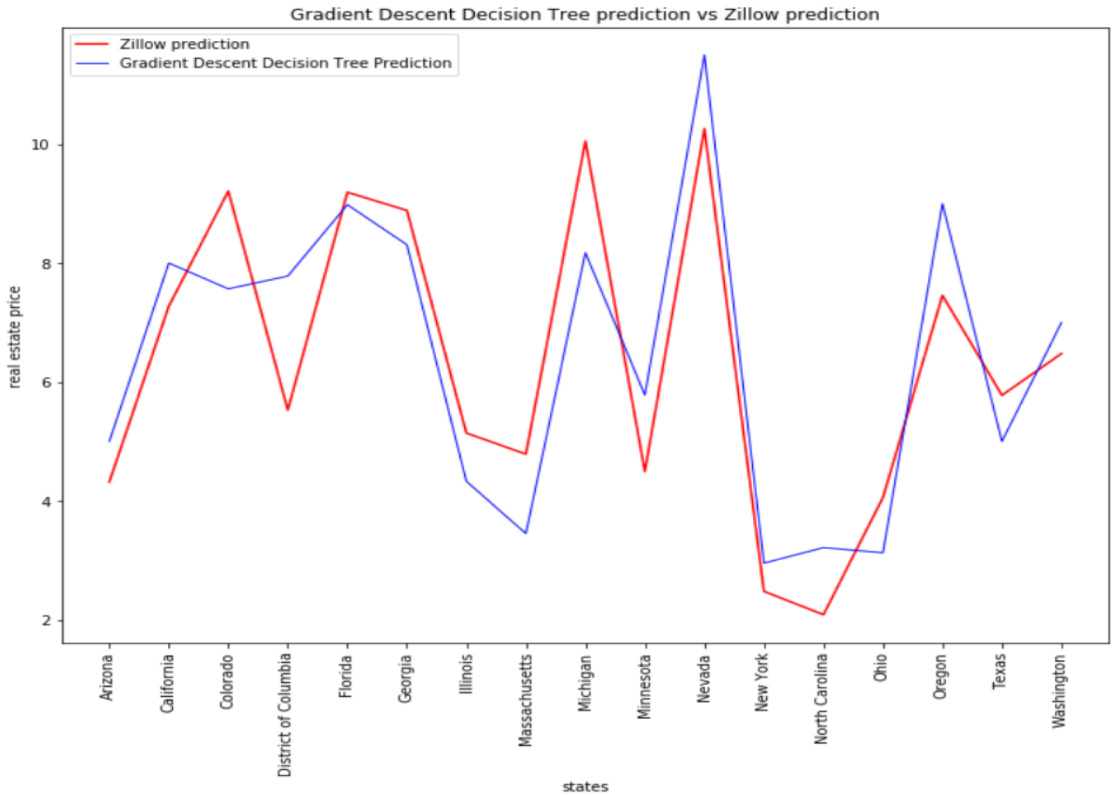
- Generate the statewide prediction plot, and can come to the conclusion that not all of them are alpha factors. Then we rank the risk factors by significance score, and thus obtain the potential alpha factors
- These include: economic/demographic factor such as Residence Pop, Violent Crime rate and Homeownership Rate, and “distressed proxy factors” such as real GDP and Median Household Income

State Level Model - OLS model w/ Momentum factor

- Momentum has risk premium, i.e., has predictive power, because investors will tend to buy house that already looks expensive
- Check the alpha factors individually to see if they has momentum
- Momentum of Median Household Income and real GDP has predictive power
- 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







State Level Model
- Xgboost: Gradient Descent Decision Tree

Main problem: Lack time series data for many counties

County Level Prediction

Exploratory Data Analysis

- **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**

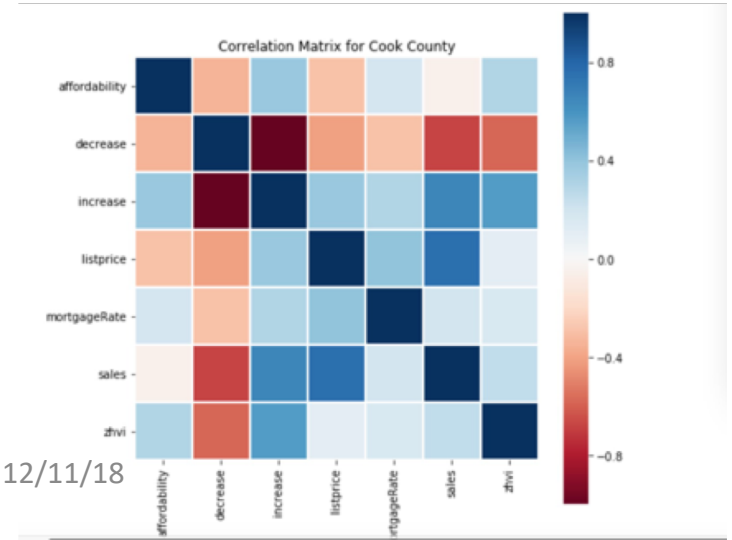
Factor Pool

We use monthly data from 2011-3 to 2018-10, since list price and sales are only available from 2011-3

- **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)
- **Affordability:** Zillow's index for mortgage affordability. Similar to price-to-income ratio. 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.
- **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.
- **Mortgage rate:** Weekly quoted mortgage rate. We calculate the mean for each month as mortgage rate factor.

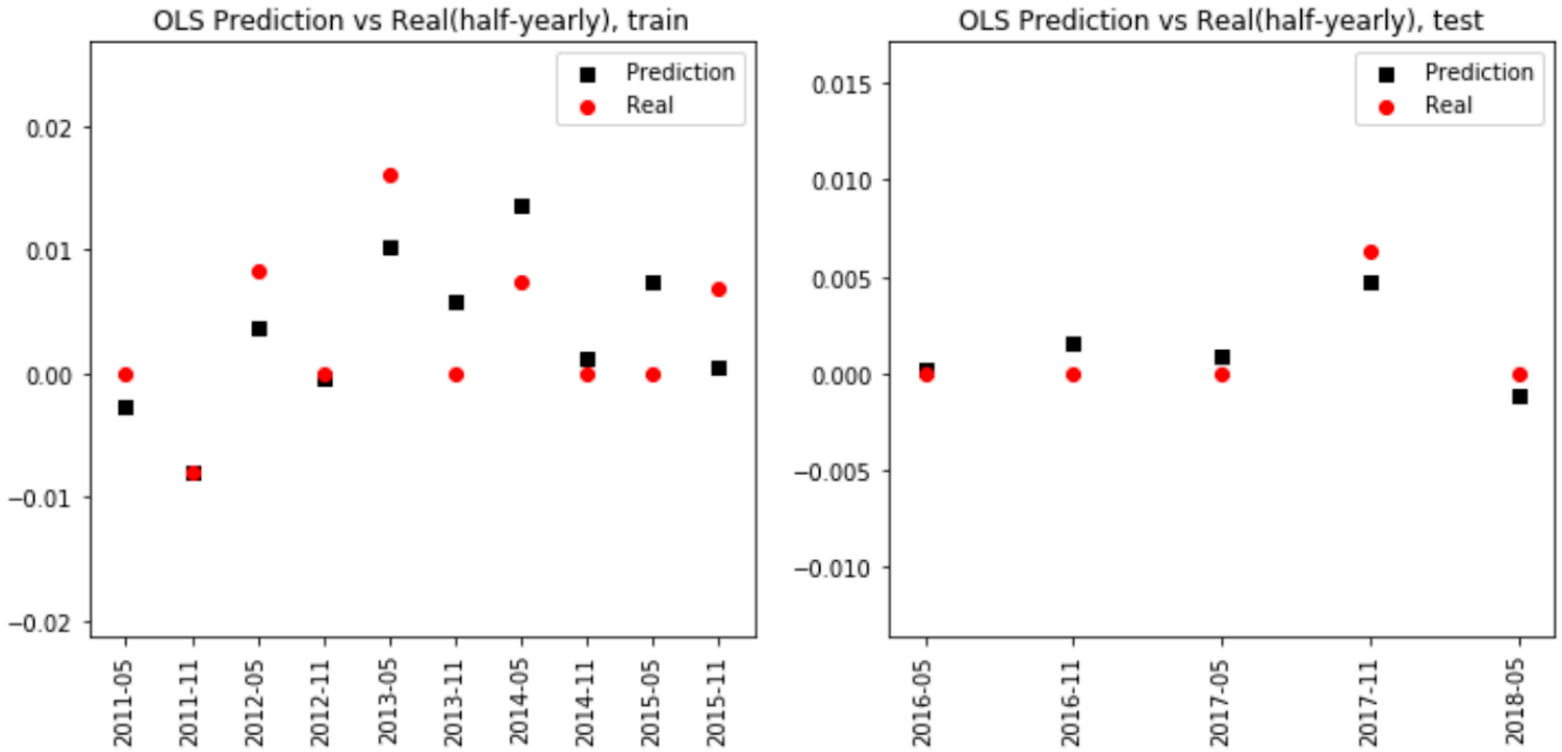
Factor Matrix and correlation

	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	



Prediction

Benchmark: OLS



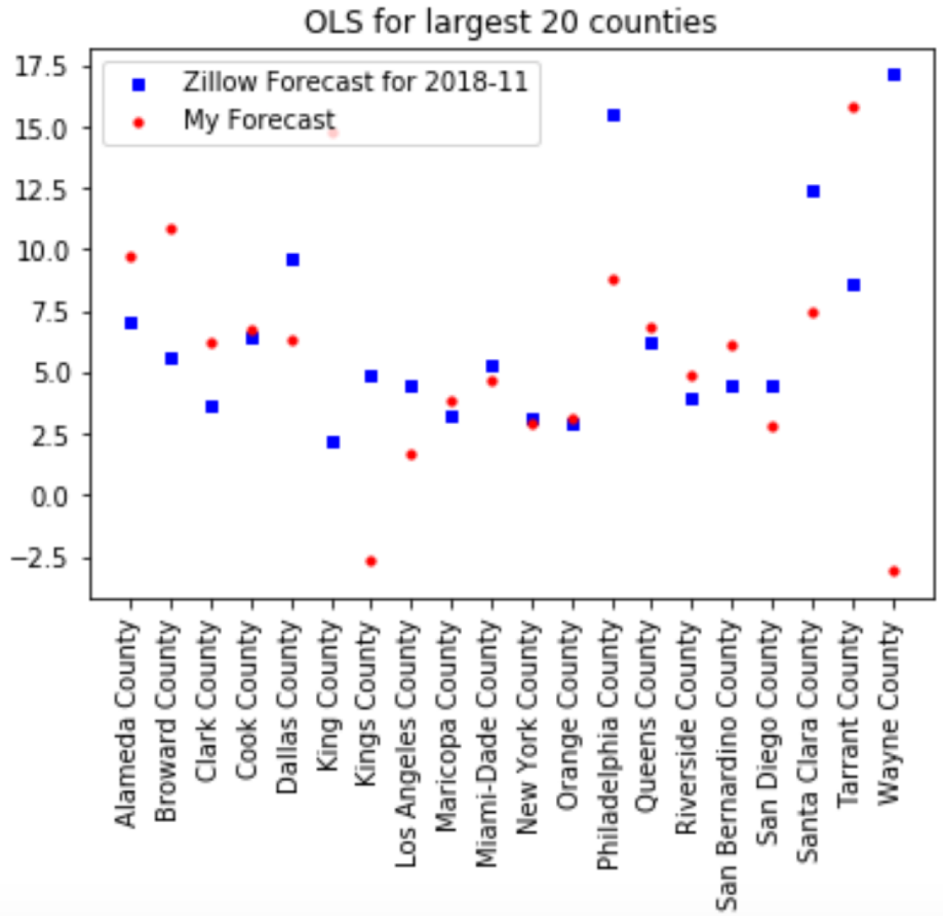
Prediction for cook county

corr_train(prediction, real) = 0.6966946

corr_test(prediction, real) = 0.88670395

Prediction

Benchmark: OLS

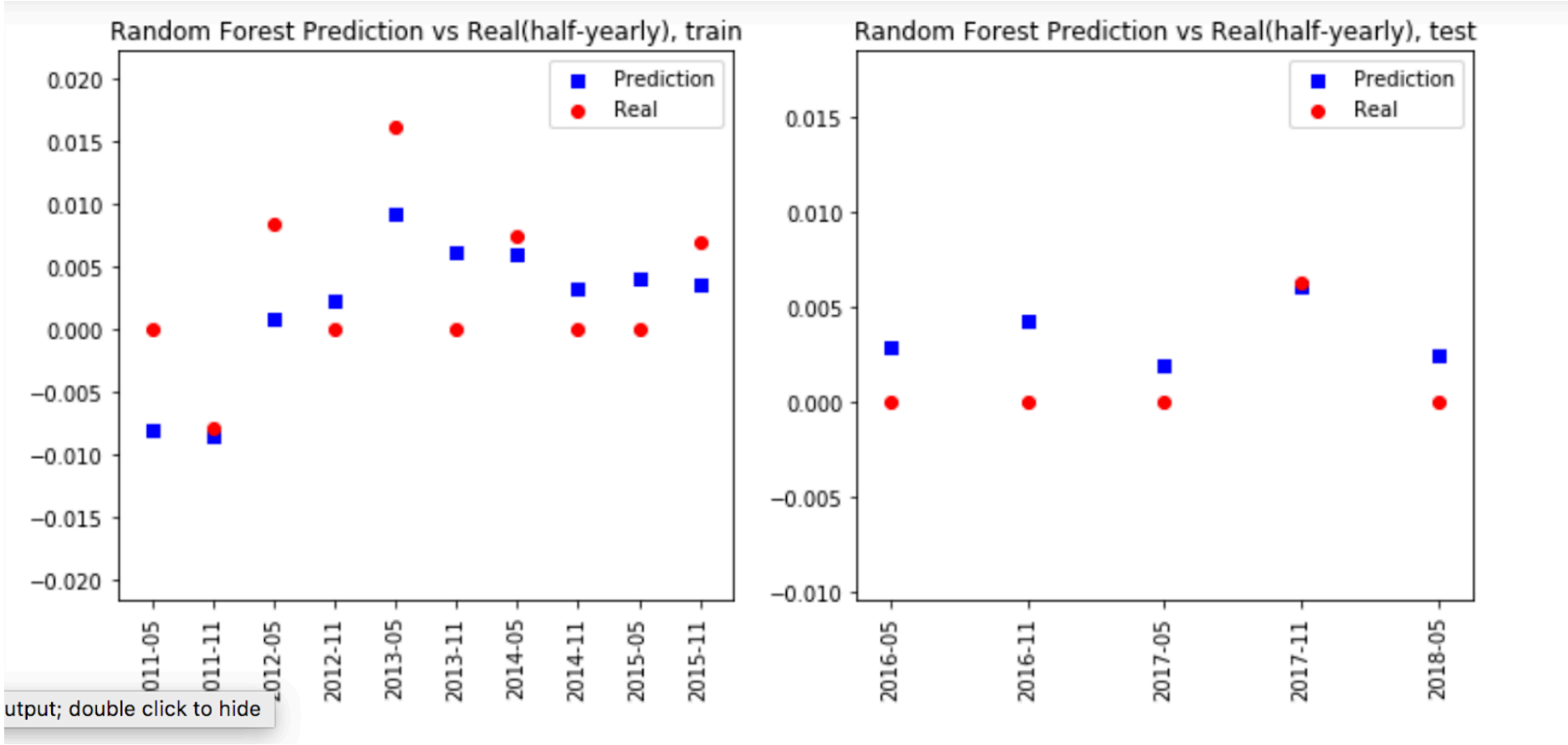


Prediction for top 20 size counties

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

Prediction

Random Forest



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

Conclusion

- **Seasonality is more important than other features**
- **Lag-ZHVI has predict power to Case- Shiller**
- **Xgboost performs very well for all the state-level**
- **Distressed proxy factors has momentum**

Thank you!