MSDS 604 Time Series Analysis Final Project

Fan Li, Qiaochu (Tina) Liu, Yuefan Wang

Data Description

The Zillow dataset (modified) recorded Feb 2008-Dec 2015 monthly median sold price for housing in California, Feb 2008-Dec 2016 monthly median mortgage rate, and Feb 2008-Dec 2016 monthly unemployment rate.

Objective

Explore the dataset and use time series analysis to find model to forecast the monthly median sold price for Jan-Dec 2016.

Dataset Split

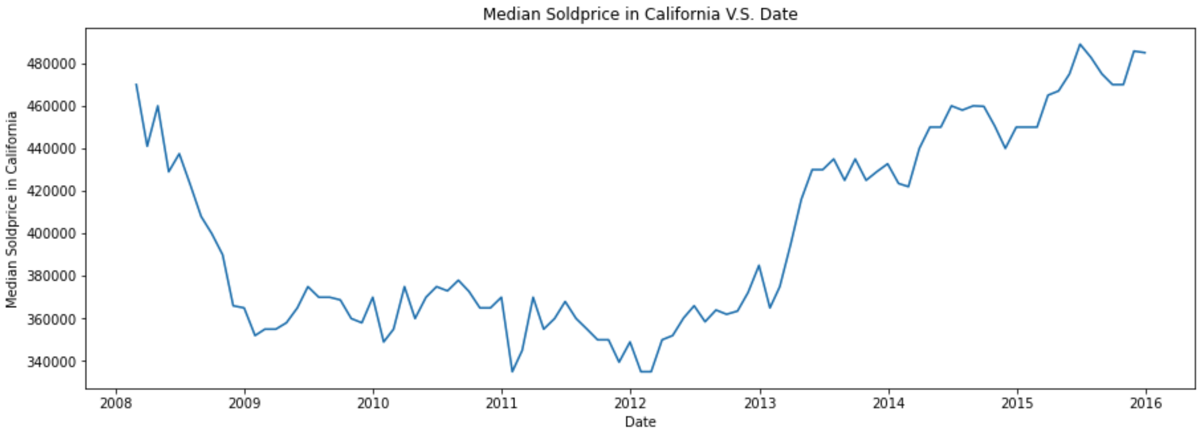
Train set included data from 2008-02-29 to 2015-12-31.

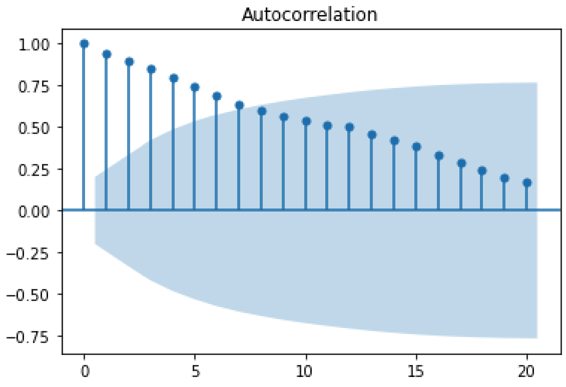
Validation set included data from 2015-01-31 to 31 - 2015-12-31.

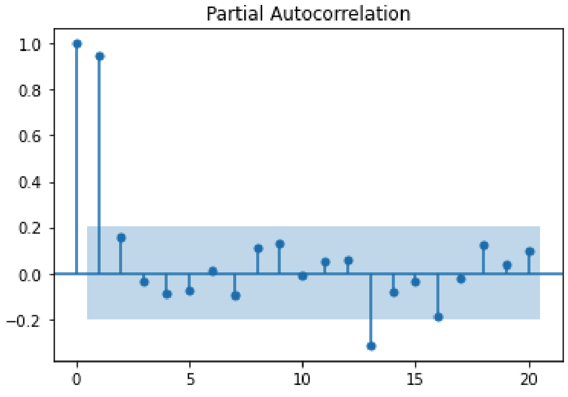
Test set included data from 2016-01-31 to 2016-12-31.

Initial Exploration

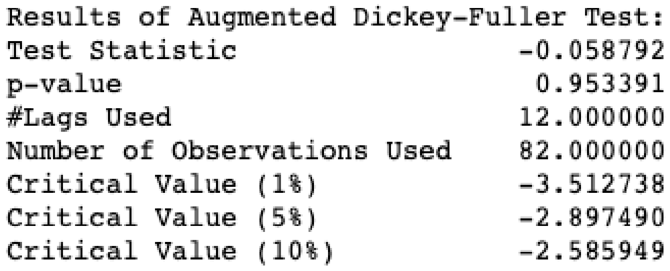
For time series analysis, we started by plotting time series, ACF, and PACF plots. Intuitively we could see a clear trend over the years. There was no obvious seasonality. However, we were still open to explore potential seasonality that might not be obvious in the plots.







We wanted to check if the orginal target varible time series data was stationary, which could deteremine whether we needed to difference the data in order to fit time series model.

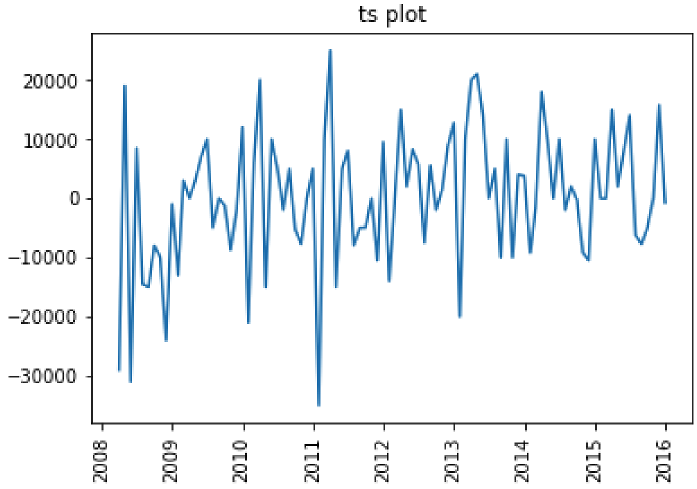


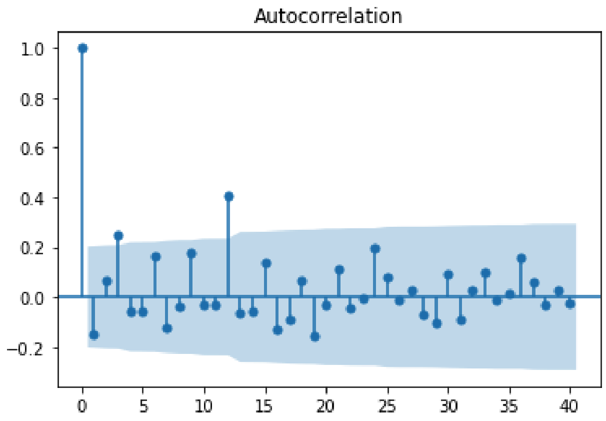
```
Results of Augmented Dickey-Fuller Test:
Test Statistic                  -0.058792
p-value                          0.953391
#Lags Used                      12.000000
Number of Observations Used     82.000000
Critical Value (1%)             -3.512738
Critical Value (5%)             -2.897490
Critical Value (10%)            -2.585949
```

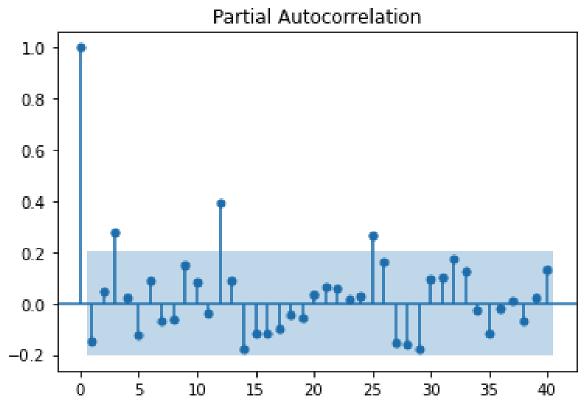
The p-value we derived from the Augmented Dickey–Fuller test was higher than 0.05, which suggested that the original target variable time series was not stationary.

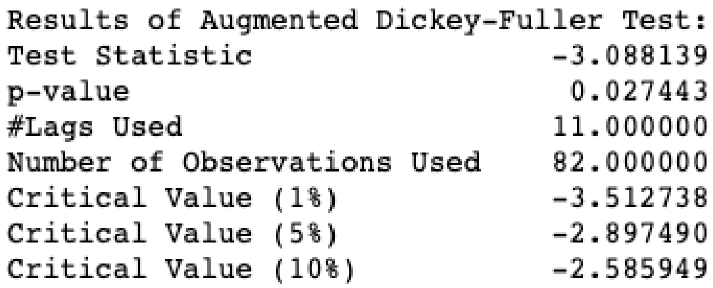
Univariable Model Exploration

From the previous exploration, we knew that we needed to difference the original time series in order to fit a model. We started with difference the data once.





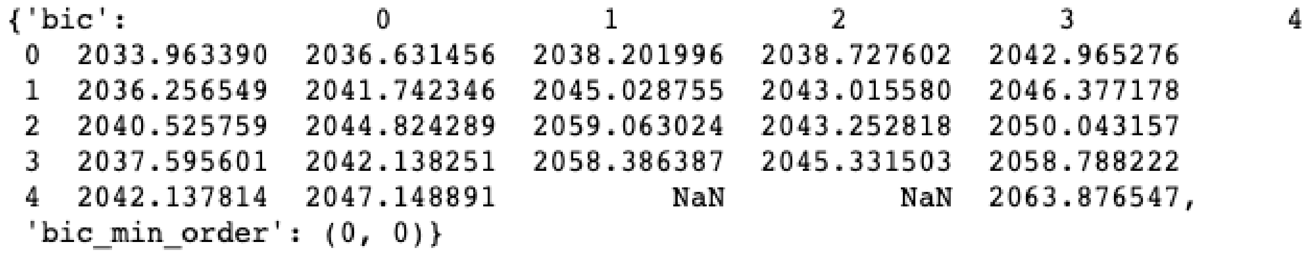




```
Results of Augmented Dickey-Fuller Test:
Test Statistic                 -3.088139
p-value                         0.027443
#Lags Used                     11.000000
Number of Observations Used    82.000000
Critical Value (1%)            -3.512738
Critical Value (5%)            -2.897490
Critical Value (10%)           -2.585949
```

From the plot and test results, we could conclude that one-differenced data was stationary. Therefore, we can start to find model candidates to fit univariable model. Since there was no obvious seasonality from the plots, we would only consider ARIMA model for univariable time series model selection.

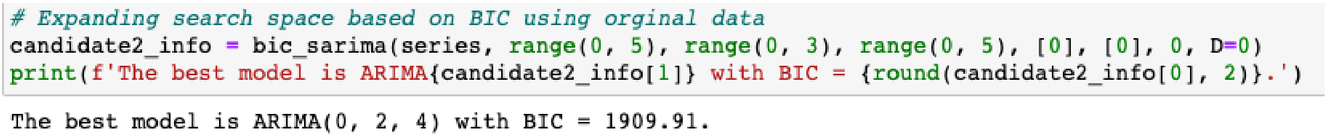
Firstly, we wanted to find the best ARIMA model candidate for one-differenced data based on BIC.



```
{'bic':            0            1            2            3            4
 0  2033.963390  2036.631456  2038.201996  2038.727602  2042.965276
 1  2036.256549  2041.742346  2045.028755  2043.015580  2046.377178
 2  2040.525759  2044.824289  2059.063024  2043.252818  2050.043157
 3  2037.595601  2042.138251  2058.386387  2045.331503  2058.788222
 4  2042.137814  2047.148891          NaN          NaN  2063.876547,
 'bic_min_order': (0, 0)}
```

We found the first ARIMA model candidate was ARIMA(0,1,0).

Secondly, we expanded the range for p, d and q to 0 - 4, 0 - 2 and 0 - 4 to do a wider grid search to find the best model using the original data set based on BIC. We found the following result:

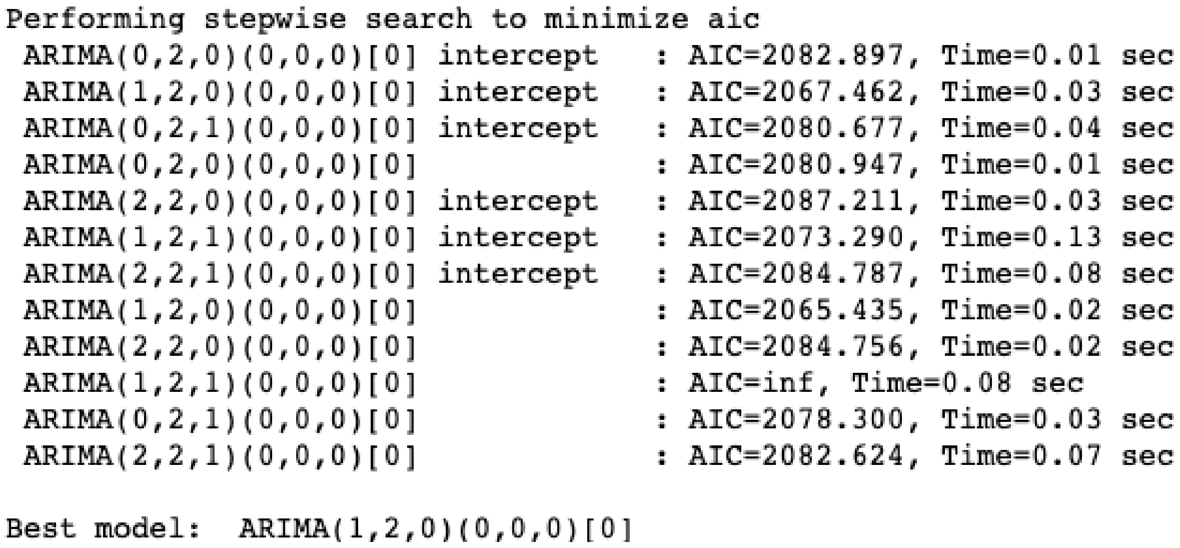


```
# Expanding search space based on BIC using orginal data
candidate2_info = bic_sarima(series, range(0, 5), range(0, 3), range(0, 5), [0], [0], 0, D=0)
print(f'The best model is ARIMA{candidate2_info[1]} with BIC = {round(candidate2_info[0], 2)}.')
```

```
The best model is ARIMA(0, 2, 4) with BIC = 1909.91.
```

Our second ARIMA model candidate was ARIMA(0,2,4).

Lastly, we used auto_arima function to derive last model candidate based on AIC.

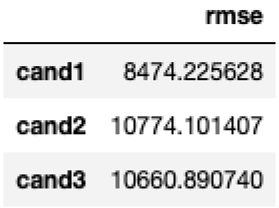


```
Performing stepwise search to minimize aic
 ARIMA(0,2,0)(0,0,0)[0] intercept   : AIC=2082.897, Time=0.01 sec
 ARIMA(1,2,0)(0,0,0)[0] intercept   : AIC=2067.462, Time=0.03 sec
 ARIMA(0,2,1)(0,0,0)[0] intercept   : AIC=2080.677, Time=0.04 sec
 ARIMA(0,2,0)(0,0,0)[0]             : AIC=2080.947, Time=0.01 sec
 ARIMA(2,2,0)(0,0,0)[0] intercept   : AIC=2087.211, Time=0.03 sec
 ARIMA(1,2,1)(0,0,0)[0] intercept   : AIC=2073.290, Time=0.13 sec
 ARIMA(2,2,1)(0,0,0)[0] intercept   : AIC=2084.787, Time=0.08 sec
 ARIMA(1,2,0)(0,0,0)[0]             : AIC=2065.435, Time=0.02 sec
 ARIMA(2,2,0)(0,0,0)[0]             : AIC=2084.756, Time=0.02 sec
 ARIMA(1,2,1)(0,0,0)[0]             : AIC=inf, Time=0.08 sec
 ARIMA(0,2,1)(0,0,0)[0]             : AIC=2078.300, Time=0.03 sec
 ARIMA(2,2,1)(0,0,0)[0]             : AIC=2082.624, Time=0.07 sec

Best model:  ARIMA(1,2,0)(0,0,0)[0]
```

The last ARIMA model candidate was ARIMA(1,2,0).

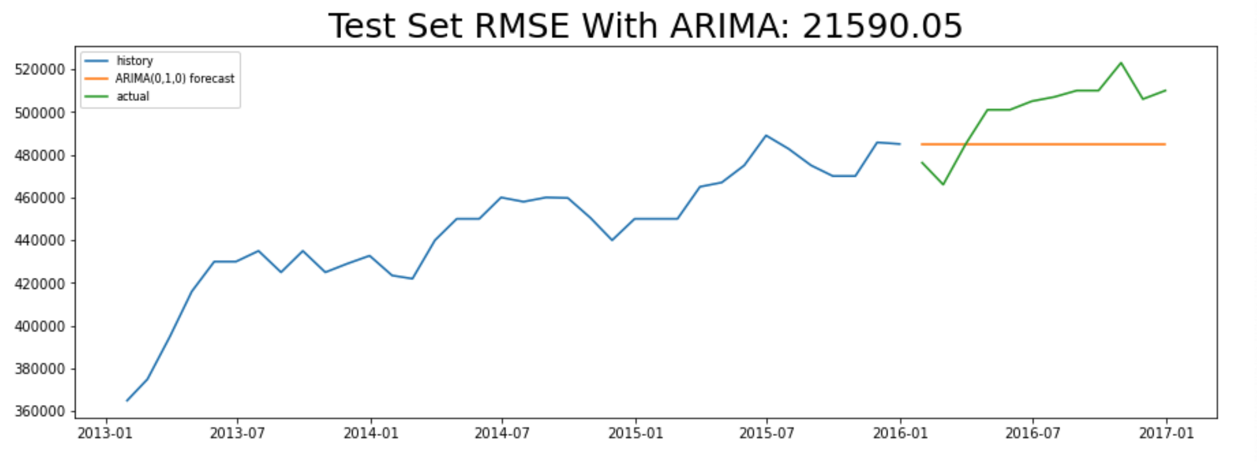
After having 3 model candidates for ARIMA model, we used cross validation with the train and validation set to determine the best ARMIA model based on RMSE.



	rmse
cand1	8474.225628
cand2	10774.101407
cand3	10660.890740

From the output above, we can concluded that first candidate ARIMA(0,1,0) has the best performance based on RMSE.

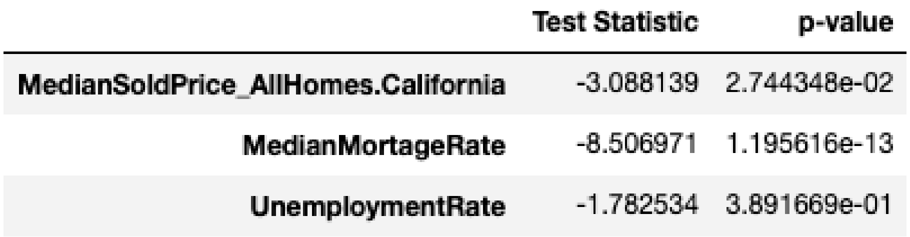
Fit the ARIMA(0,1,0) with the entire training set from 2008-02-29 to 2015-12-31 and forecast the monthly median sold price for January to December 2016.


Test Set RMSE With ARIMA: 21590.05

We concluded that the univariable model may have a good performance to fit the training set, but its prediction power was poor as it quickly fell to the value of last observation in the training set. We were going to explore more about the data set including some exogenous variables using multivariable time series models.

Multivariable Model Exploration

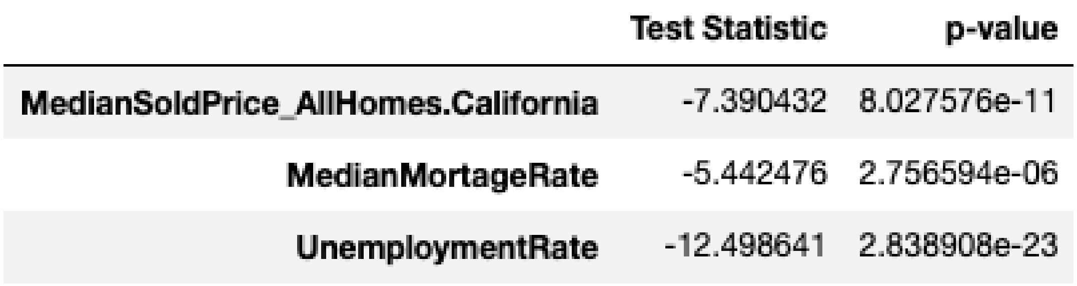
At first, we wanted to consider the VAR model. It required data to be stationary, thus the first task was to detrend the data. From the previous discussion, we need to at least difference the data once.



	Test Statistic	p-value
MedianSoldPrice_AllHomes.California	-3.088139	2.744348e-02
MedianMortageRate	-8.506971	1.195616e-13
UnemploymentRate	-1.782534	3.891669e-01

`UnemploymentRate` is still not stationary

After differencing the entire dataset once, “UnemploymentRate” was still not stationary. We would difference it one more time.

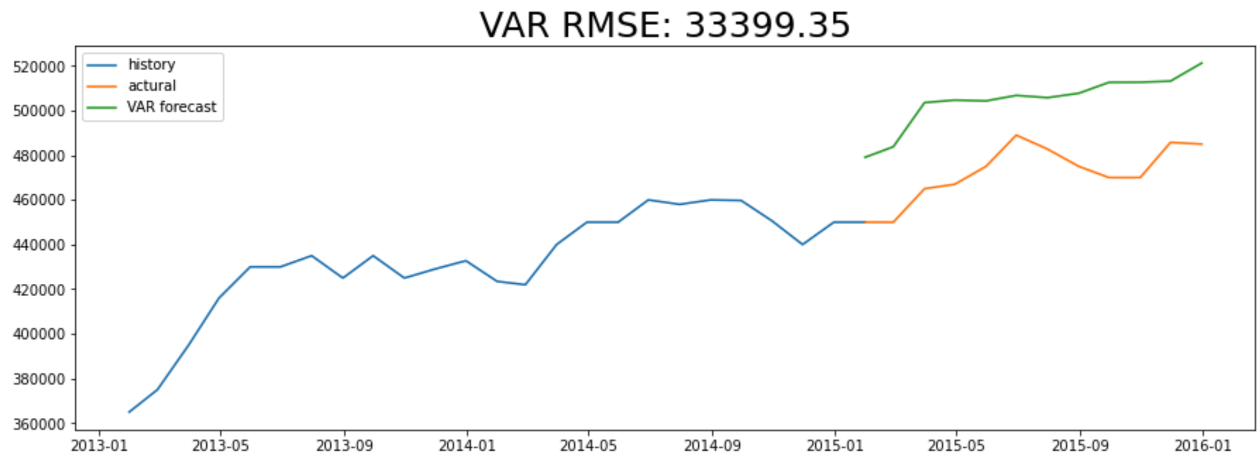


	Test Statistic	p-value
MedianSoldPrice_AllHomes.California	-7.390432	8.027576e-11
MedianMortageRate	-5.442476	2.756594e-06
UnemploymentRate	-12.498641	2.838908e-23

Differencing twice all variables reach stationary

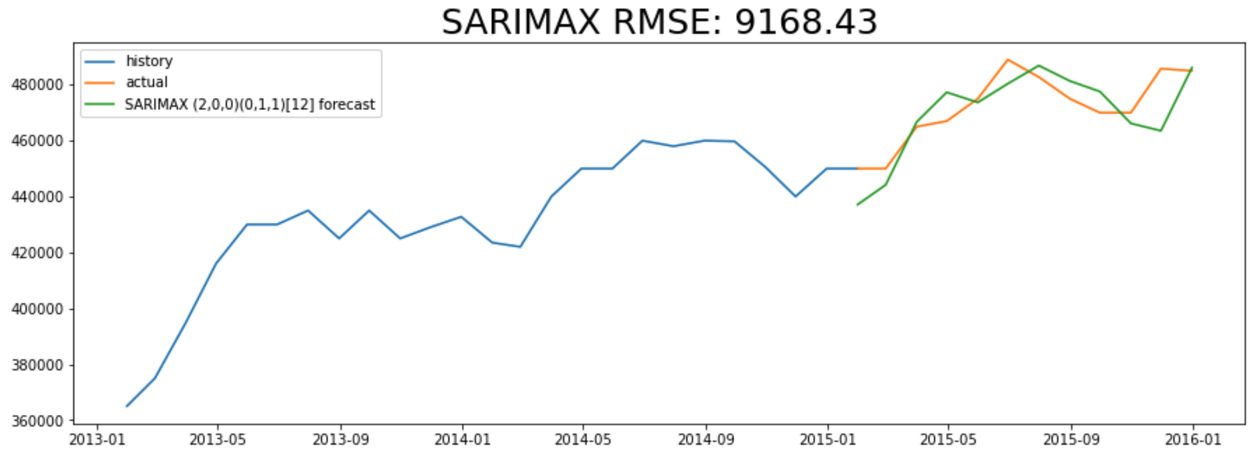
After differencing the original data twice, we fitted the VAR model based on AIC. The coefficient of the model we got was p = 11.

We plotted the true validation set against prediction and computed its RMSE after inverting the data back to the data prior to the differences. As we can see here, VAR captured the general trend but it’s very biased. Overall RMSE is almost 33400.



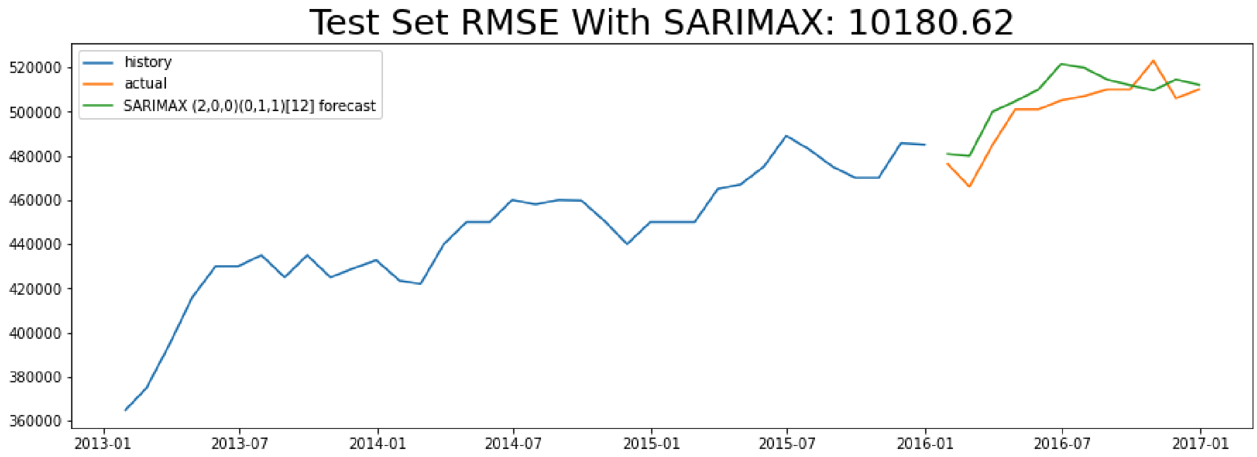
All the features in the data set itself should to be endogenous and influence each other, thus the VAR model should make the most sense, but RMSE was not promising. We decided to explore the SARIMAX model.

For the SARIMAX part, we first normalize the features for better prediction, later we would use the same train set scaler to normalize and transform the test set as well. We used auto SARIMAX search with a seasonality lag m = 12 (12 months in a year), set columns ‘MedianSoldPrice_AllHomes.California’ as the endogenous and rest as exogenous and open to other hyper-parameters tuning. The final model had an order of (2,0,0)(0,1,1)[12] and a very decent RMSE score when we fitted the same period validation set as VAR model.



Between those two models, The SARIMAX outperformed the VAR model based on the plot and RMSE. We would use SARIMAX as the final multivariate model.

We plotted the fitted SARIMAX with the entire training set to see its prediction power on the test set.



The SARIMAX model also had a good performance for forecasting the test set.

Conclusion

We explored both univariable and multivariable time series model using consistent training, validation and test set. We found that multivariable had a better performance to forecast monthly median sell price from January to December 2016 based on both prediction graphs and RMSE. As SARIMAX model did consider other variables, (MedianMortageRate and UnemploymentRate) in the dataset to forecast the house sold price and had better performance than ARIMA model, it suggested that mortgage rate and unemployment rate might have some effect on the median house sold price in 2016.