Time Series Model for California Zillow Data 2008-2016

Roger Ren, Zhipeng Hong, Kyril Panfilov

Dataset and Goals

For this analysis, the dataset contains monthly median sold price in California, monthly median mortgage rate, and monthly unemployment rate for the time period February 2008 through December 2016.

Field	**Definition**	**Data Type**
Date	Last date of each month	date
median_price	California monthly median home sell price	float
median_rate	Monthly median mortgage rate	float
unemploy_rate	Monthly unemployment rate	float

The goal of the project is to forecast median California house prices. The main focus will be to use the median price history to model at time series that will be used to forecast future prices. We will also explore the potential of including mortgage rate and unemployment rate in multivariate models to see if it helps us make better predictions of median house prices.

The 2016 data is cut out for testing of the final model's performance, so that the candidate models were trained and evaluated on data from 2008 through 2015. This allows us to evaluate the generalized performance of the final model on future unseen data.

Methods

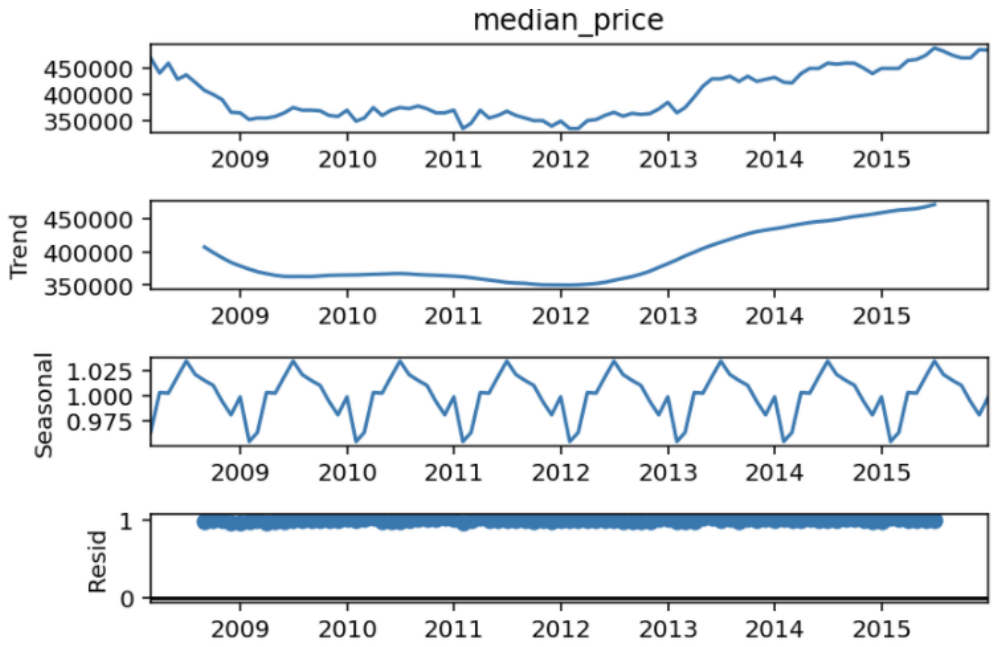
When modeling a time series, it is important to consider multiple modeling methods, selecting the best performing candidate models from each method, and then select a final model by comparing the evaluation of each candidate model.

Setup

First, we split the 2008-2015 data into training and validation using a 80/20 sequential split, so that we can train our candidates on the training data and forecast predictions to be evaluated on the validation data.

We chose the Root Mean Squared Error (RMSE) metric in order to evaluate the forecasting performance of the models. RMSE is a measure of how much our forecasts deviate from the actual validation values. RSME is helpful to tell us which models are performing better, where a higher RMSE is worse than a lower RMSE.

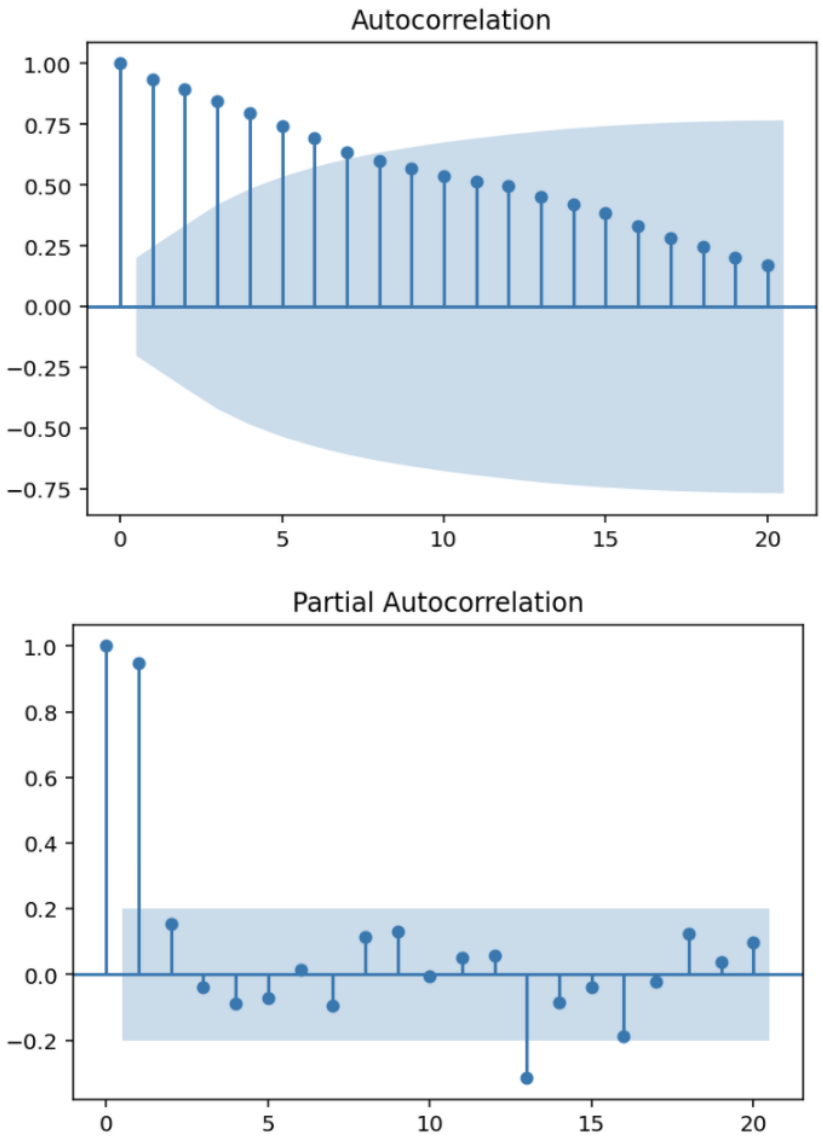
Before applying potential models to our training data, we made a few visualizations of the seasonal decomposition to understand the seasonality and trend.



This decomposition shows a downward and then upward trend, meaning we may need to apply differencing for models that require stationary time series. It also shows us a consistent 12 month seasonality, which we can apply to our SARIMA models. Lastly, the above plot assumes a multiplicative seasonality which is confirmed by the residual plot.

Since the seasonal decomposition plot seems to show some trend, we used an ADF test to see if any differencing was required. Without differencing we found the ADF test gave us a p-value of 0.95, which is much higher than any standard significance level, meaning the time series is not stationary and would require some differencing. After differencing the data once, our ADF test p-value improved to 0.02, which is below our selected 0.05 significance level. This signaled to us that we should look to difference our data when using the different modeling approaches in the next step.

Before continuing with complex model choices such as ARIMA and SARIMA, we plotted the ACF (autocorrelation) and PACF (partial autocorrelation) plots to see if we can identify any basic AR (autoregressive) and/or MA (moving average) order processes visually.



Since the ACF plot is decaying and the PACF plot shuts off after 2. This indicates AR(2) will be a selection to consider when applying ARIMA and SARIMA models to the training set.

Modeling

For modeling, we chose to try ARIMA, ETS, SARIMA, and SARIMAX models. The ARIMA stands for Autoregressive Integrated Moving Average, which essentially uses the linear relationship, past shocks (unexpected events), and general trend to model and predict future values. We tested a variety of ARIMA models using a grid search, and selected the ARIMA model with the lowest RMSE on the validation set. Based on our testing in the previous step, we made sure to include differencing and AR(2) in the grid search. This is a univariate model that only uses the history of median California house prices to model and forecast future results. The

selected ARIMA model (order=(1,1,2)) gave us an RMSE of 14,257 with which we used to compare to the other models.

ETS stands for Exponential Smoothing, which is a special case of ARIMA and SARIMA models that estimates the same way as these models, but predictions are made using exponentially decaying weights instead of updating the weights based on time. ETS is also a univariate modeling technique. After testing the multiple variations of the ETS models, we found that it was best to use a multiplicative trend and seasonality, which gave us an RMSE score of 10,176 and seemed to outperform the ARIMA model based on the validation data.

Since we had observed seasonality in our seasonal decomposition, we also tested the SARIMA model, which is an ARIMA model that adds a seasonality component. Based on our observations in the setup step, we performed a grid search of potential parameters and selected the model with the lowest AIC, which used ARIMA order (1,1,0), seasonality component (1,1,3), and a 12 lag for the 12 month seasonality. This model had an improved RMSE of 5,744 on the validation set.

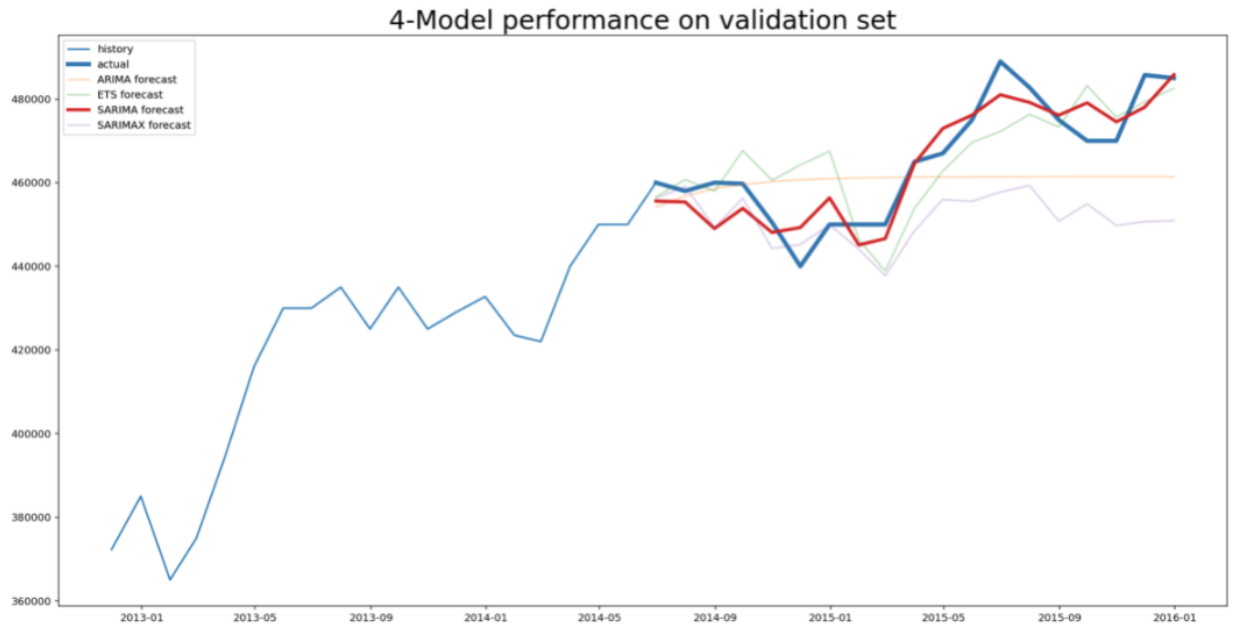
The final model we tested was the SARIMAX, which is a version of ARIMA that allowed us to add seasonality parameters and use multivariate data. Since our goal is to forecast median house prices, we chose the median_price column as our endogenous variable and the rest as exogenous. Then, we performed a grid search and selected the model with the best AIC score. The evaluation on the validation set resulted in an RMSE score of 20,751 which is worse than the univariate SARIMA model results.

Findings

To summarize the results, the SARIMA model had the lowest RMSE score on the validation data.

Model	**Best Parameters**	**Best RMSE**
ARIMA	Order : (1, 1, 2)	14,257
ETS	(mul, mul)	10,176
SARIMA	Order: (1, 1, 0)(1, 1, 3)[12]	5,744
SARIMAX	Order: (4,1,0)(1,0,0)[12]	18,221

To illustrate the difference, we also plotted the forecast of our best performing model forecast on the validation data set.



Visually, we can see the red SARIMA model forecast is closest to the blue actual values from the validation data.

Forecasting

After selecting our final SARIMA model, we can evaluate how well our model will generalize on future data by testing our forecast against the test data saved from 2016. These are the predictions for median home prices in 2016 shown next to the actual values.

	predicted_median_house_price	actual_median_house_price
2016-01-31	475547.708547	476250
2016-02-29	475440.794225	466000
2016-03-31	493510.561173	485000
2016-04-30	501792.294174	501000
2016-05-31	511429.650049	501000
2016-06-30	519963.996041	505000
2016-07-31	518067.569771	507000
2016-08-31	510997.153610	510000
2016-09-30	511971.784103	510000
2016-10-31	507731.871501	523000
2016-11-30	514728.466645	506000
2016-12-31	518031.555774	510000

The resulting RMSE from this forecast is 9,112 which still performs better than the other models on the validation data. Finally, we show visually how close our model performs relative to the actual test data. Orange is the SARIMA forecast and blue is the actual values.