Time Series Forecasting Median Sold Price of Houses in California

Group Members:
Yixuan Zeng, Dominnic Chant, Shubham Thakur

1. Introduction

1.1 Description of Data

In this report, we are using Zillow's monthly Median sold price of all the houses present in California. The dataset includes the Date, MedianSoldPrice_AllHomes.California, MedianMortageRate, UnemploymentRate. 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. The dataset has in total 107 rows. We have used the historical data of median house price and exogenous variables like Median Mortgage rate, Unemployment rate to forecast the monthly median sold price for Jan-Dec 2016. We found that all the variables are numerical variables except Date which was present in YYYY-MM-DD format.

1.2 Research Questions and Approach

While beginning the project we wanted to see if the series is in the right shape to be modeled and if there are any exogenous variables that are significantly impacting the forecast. Keeping that in mind, we had the following research questions:

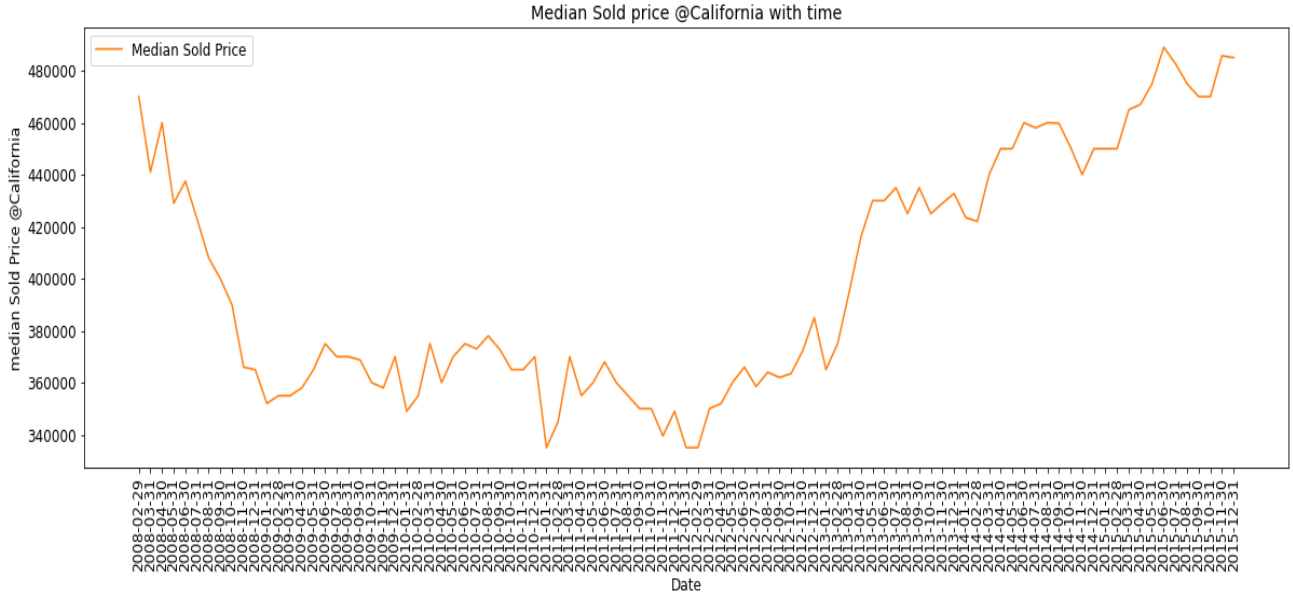
- Is the Series Stationary?
- What should be the appropriate seasonal period for the data?
- Do the exogenous variables have any impact on the Median Sold price of the houses?

In order to answer this question, we conducted an in-depth analysis using all the variables in the dataset. For the first question we used a stationarity test like Dicky Fuller to figure out if the present data is stationary or not. Then we checked the seasonal decomposition of the time series to get the understanding of the seasonal period that we must be considering while modeling time series. We then used multiple time series modeling approaches to get the best candidate models. Finally, we performed cross-validation to get the best model using RMSE as the evaluation metric. After getting the best model we tested it on the test set.

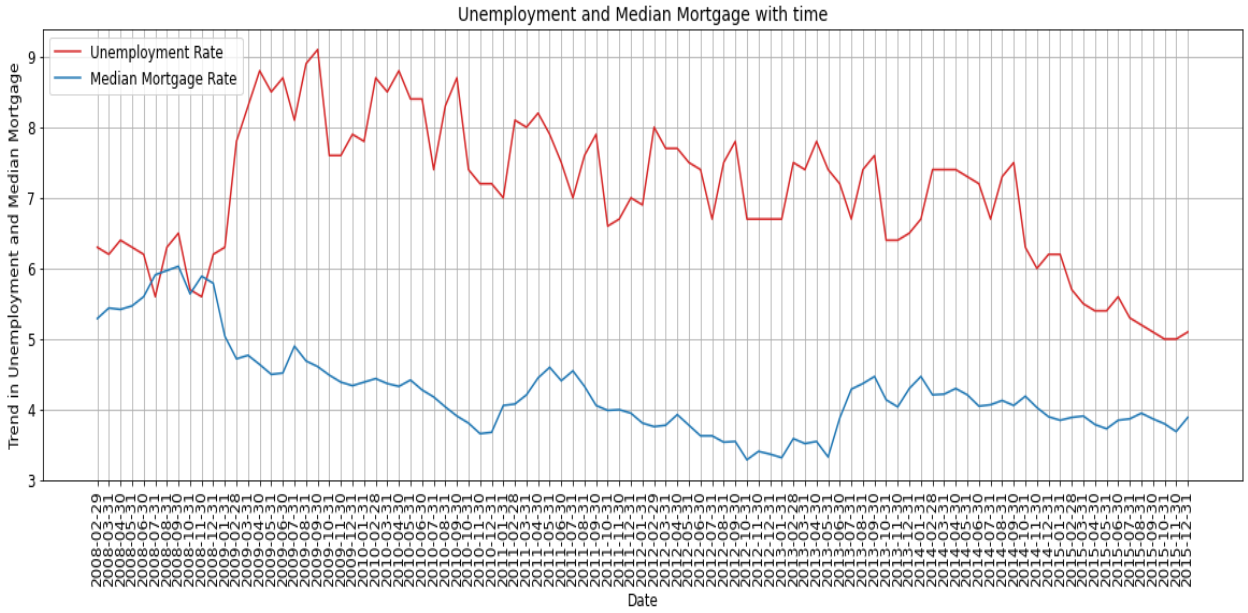
1.3 Exploratory Data Analysis

Our first step of data analysis is to make sure all the variables are in the right format. Then we checked the missing values in the data set. We found that between Feb 2008- Dec 2015 there were none which for our case is the training data. We considered data after 2015 as our test set.

As a part of visualization, we plotted the median sold price of all the houses in California. This gave us some idea of Dickey-Fuller the seasonal period for our data.

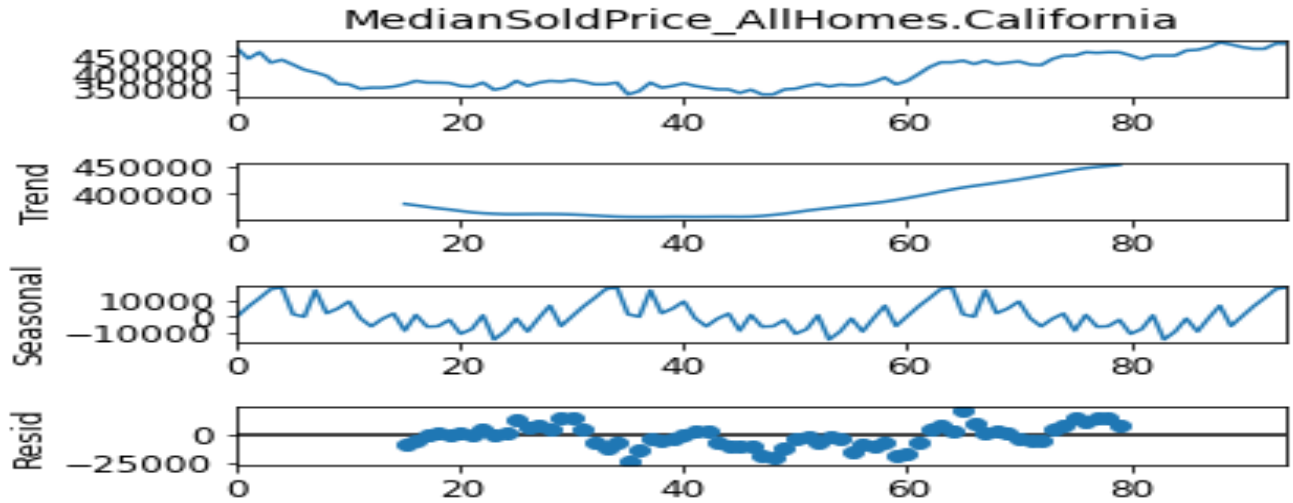


From the plot, we can see that there is some pattern after 12 months, i.e annually. However, for some years there is some sudden drop and sudden jump in the house prices. It could be possible that these are linked to the Unemployment rate or Mean Mortgage rate. As when the unemployment rate is skyrocketing we can say the cost of houses might suddenly drop. Thus we checked the pattern in the following exogenous variables.



As supported by our hypothesis, we can see there that the unemployment rate and house prices have in general a negative relationship. So thus the house prices are increasing when the unemployment rate is decreasing. Similarly, on a broad scale, the mean mortgage rate has a positive correlation with house prices.

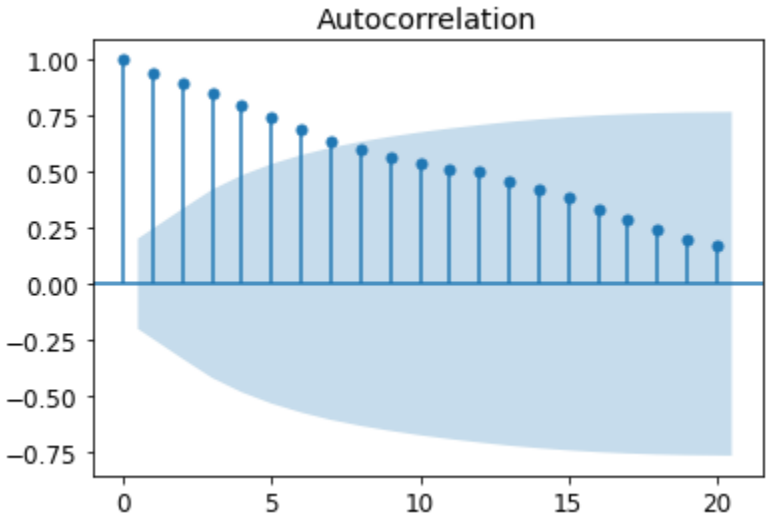
We then decomposed the time series in trend, seasonality, and errors. This gave us a rough idea about the general trend, and seasonality in the house prices. From the below graph we can say the general trend is decreasing for some time and then it started to increase. Also, we can see the seasonality pattern that is repeating after some time. The mean of residuals is also zero for the time series.



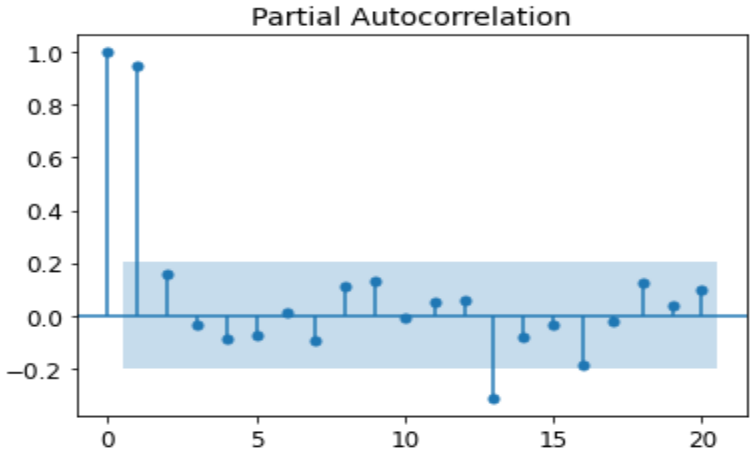
2. Model Selection

2.1 Test of Stationarity, ACF/PACF, and ARIMA Model

Initially, we checked the Autocorrelation and Partial Autocorrelation in the time series. Autocorrelation shows the correlation between different points in time series. And Partial autocorrelation gives the correlation between two points given all the other points are constant. It looks like the ACF plot is decaying and there is no shut-off after some time. So this is the MA(0) process.



In the PACF plot, there is shut off after h = 1 which suggests that this is an AR(1) process.



In total, we can say the series can be modeled using the ARMA(1,0) process. However, to use the general models from the ARIMA family, our time series needs to be stationary. We used the Dickey-Fuller test to check the stationarity of the time series. Dickey fuller test assumes:

H0: There is a unit root in an AR model, which implies that the data series is not stationary
H1: The data series is stationary.

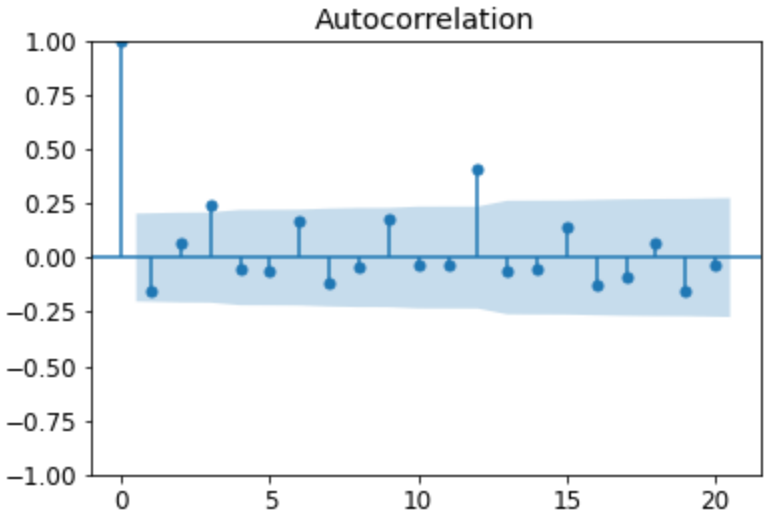
ADF test Results:

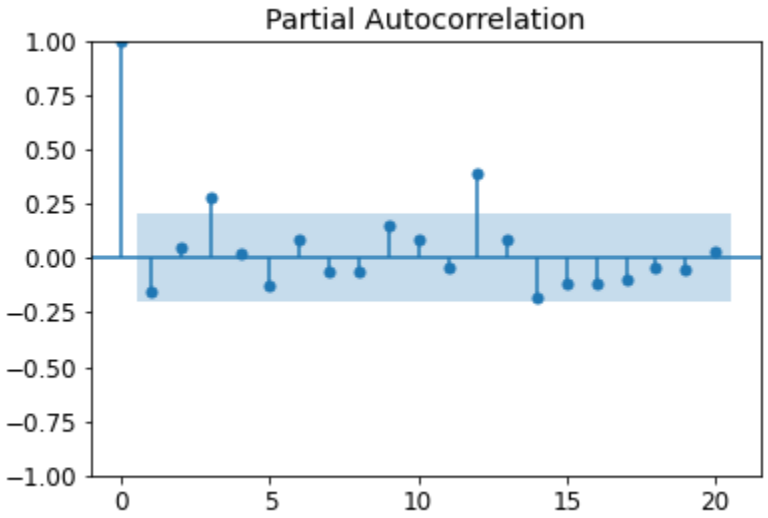
Difference Times	**p-value**	**stationary**
0	0.953	No
1	0.027	Yes
2	0.012	Yes

When we use the original data without a difference, the p-value is 0.953 which is higher than 0.05 which suggests that the series is not stationary. Thus we did the difference of the series once and checked if it is coming out to be stationary then. We again ran the Dickey-Fuller test to check stationarity.

The p-value then came out to be 0.027 which is lower than the significance level of 0.05. This suggests that the differenced time series is stationary. This is equivalent to modeling an ARIMA(1,1,0) series. Therefore our first ARIMA candidate model is ARIMA (1,1,0).

After we differentiate and conduct an ADF test for the difference data, we check the ACF plot and PACF plot of the data we differenced once.





Since both the ACF plot and PACF plot shut off at the value of 1, the ARIMA candidate model 2 is (0,1,0).

Furthermore, since the data is stationary when we difference the data once and twice, we conduct the order search using BIC as criteria for the ARIMA model with the d-value equals to 1 and 2 respectively. When d=1, the order search gives us the same order as ARIMA candidate 1. When d=2, the search suggests that the best model is (0,2,1), which is our third ARIMA candidate model.

After having 1-step cross-validation, here we have our result of ARIMA models:

ARIMA order	RMSE	MAE	MAPE
(1,1,0)	8230.734	6302.597	0.0135
(0,1,0)	8161.608	6236.842	0.0133
(0,2,1)	10323.724	8492.500	0.0182

All three matrices suggest that the second model is the best ARIMA model.

2.2 SARIMA model

From the ADF test above, we see when d = 1, the data becomes stationary. Therefore for the SARIMA model. We also take BIC order search with d=1, and it suggests that the best model is SARIMA (0, 2, 2), (1, 1, 2, 12). This is our SARIMA candidate 1 model.

Then we conduct another order search using the auto_arima function with no constraint in d value, and it gives back the best SARIMA model order is (1, 2, 0), (0, 1, 0, 12), which is our second SARIMA candidate model.

Similar to the ARIMA model, we also have a SARIMA model evaluation using the 1-step cross-validation. The result is as below:

SARIMA order	RMSE	MAE	MAPE
(0, 2, 2), (1, 1, 2, 12)	12217.640	9467.574	0.0202
(1, 2, 0), (0, 1, 0, 12)	15094.505	12837.486	0.0276

According to the result above, the best SARIMA model is the one with order (0, 2, 2), (1, 1, 2, 12).

2.3 Exponential Smoothing (ETS)

The next model we are considering is Exponential Smoothing. Since the orders of ETS models are slightly different from the models from the (S)ARIMA family, which include

trend (multiplicative, additive, and None) and seasonal (multiplicative and additive). Our strategy is to try all of the combinations of these two orders and find which is our best model.

Again, we use 1-step cross-validation for model selection in Exponential Smoothing and pick the one with the lowest matrix score. The result table is as follows:

ETS Parameters	**RMSE**	**MAE**	**MAPE**
(add, add)	7868.555	5455.137	0.0117
(mul,add)	8561.613	5894.934	0.0276
(None, add)	8356.415	5776.018	0.0123
(add, mul)	127829.414	34930.764	0.0786
(mul, mul)	381264.818	96828.981	0.2147
(None, mul)	115952.117	32076.345	0.0721

Therefore, our best Exponential Smoothing model is the one with both trend and seasonal order “additive”.

2.4 Multivariate: SARIMAX

For multivariate time series models, since the data we used is not daily based, it is not suitable for us to use the Prophet model. Therefore we simply try the SARIMAX model for multivariate variables. We use median mortgage rate and unemployment as exogenous features and conduct an order search with the auto_arima function. The search return us the best SARIMAX model with order (4,1,0), (0,1,2,12).

After cross-validation evaluation, its matrix values are:

SARIMAX Order	**RMSE**	**MAE**	**MAPE**
(4,1,0), (0,1,2,12)	10486.682	7629.202	0.0164

2.5 Model Selection Conclusion

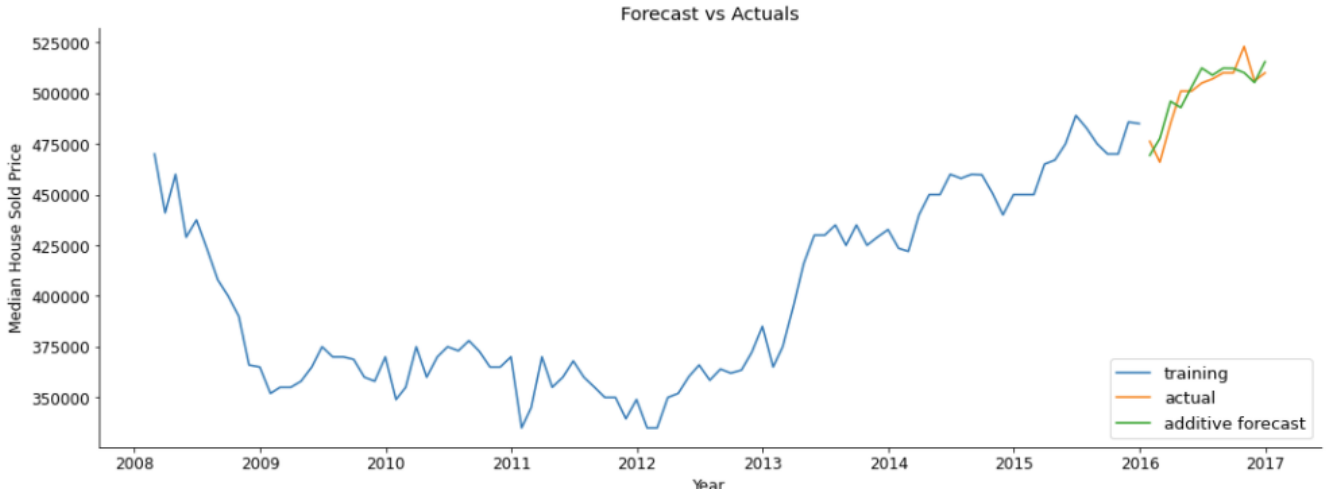
The results of our model selection are as below:

Model	Model Parameters	RMSE	MAE	MAPE
ARIMA	(1,1,0)	8230.734	6302.597	0.0135
	(0,1,0)	8161.608	6236.842	0.0133
	(0,2,1)	10323.724	8492.500	0.0182
SARIMA	(0, 2, 2), (1, 1, 2, 12)	12217.640	9467.574	0.0202
	(1, 2, 0), (0, 1, 0, 12)	15094.505	12837.486	0.0276
ETS	(add, add)	7868.555	5455.137	0.0117
	(mul,add)	8561.613	5894.934	0.0276
	(None, add)	8356.415	5776.018	0.0123
	(add, mul)	127829.414	34930.764	0.0786
	(mul, mul)	381264.818	96828.981	0.2147
	(None, mul)	115952.117	32076.345	0.0721
SARIMAX	(4,1,0), (0,1,2,12)	10486.682	7629.202	0.0164

Among all the models we have selected from the order search and its corresponding matrix values suggested in the above table, we decide to choose the ETS model with additive trend and additive seasonality since all of its cross-validation matrix scores are the lowest among the models we have evaluated.

3. Final Model Forecasts

After looking at all of the candidate models, we concluded that the ETS model with an additive order for both trend and seasonality performed the best based on the RMSE, MAPE, and MAE scores. The next step is to fit this final model on the entire train data and to make our final forecasts.



In the above table, we can see a clear upward trend of median house sold prices starting from 2012. In blue, we have the data we used to train our models on, in orange is the actual median home sold prices we are trying to predict, and lastly in green is our predictions. Our model does a decent job at forecasting the trend and seasonality of the actual data which resulted in an RMSE score of 7307.86.

4. Conclusion

The goal of this report is to showcase the best model to forecast future median house sold prices. In the process of doing so, we first did some exploratory data analysis to see if there are any glaringly obvious models we should be including in our testing. After coming up with one ARIMA model from our EDA tests, we broke down our candidate models into two types, univariate and multivariate. Within the univariate model types, we tried the (S)ARIMA family models as well as Exponential Smoothing (ETS) models for a total of eleven models. Lastly, for the multivariate case, we only chose the SARIMAX model to test because the other multivariate models did not perform that well. The final model we decided on was the ETS model with an additive order for trend and seasonality. Judging from the graph it would seem that our ETS model was indeed able to capture the trend and seasonality of the true values.