**Time Series Final Report**

By Karsten Cao, Chandan Nayak, Yangzhou Tang

# Introduction:

In this report, we explore three different methods to forecast the Median Sale Price of California houses in 2016 using historical data from February 2008 - December 2015. The time series methods that we will use to explore this time series dataset are Exponential Smoothing, SARIMAX, and Facebook's Prophet.

Exponential Smoothing is a time series analysis tool that models a trend and seasonal change. It is a specific version of the ARIMA family that focuses on the moving average of past noise with the addition of trend and seasonality. It is a fast model to use because it is restrictive in its assumptions and has few parameters to select. Because of the ease of implementation and speed, the univariate version of this model will be used as a baseline model.

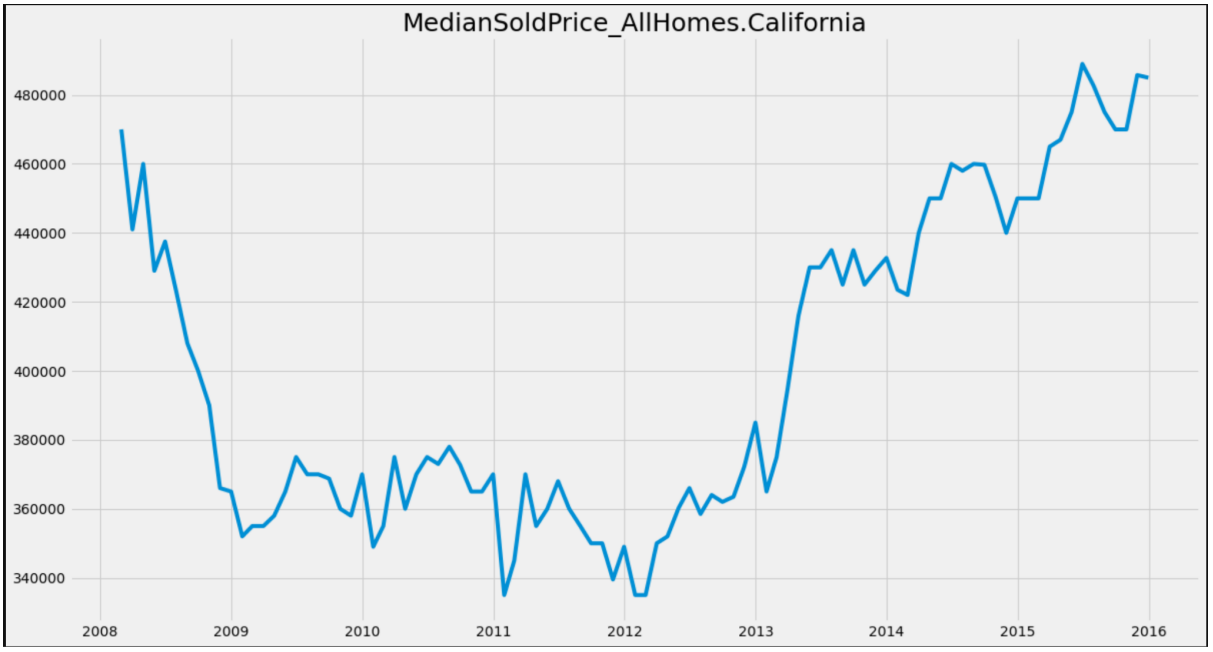
SARIMAX is a complex version of ARIMA which includes an autoregressive integrated moving average process. The 'S' in SARIMAX represents seasonality and the X represents an exogenous variable that can be considered in the model. Therefore, it requires 4 more seasonal variables than the ARIMA model. The advantage of SARIMAX is that it can take exogenous variables which will handle the potential relationship between the predicted outcomes and other variables.

Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

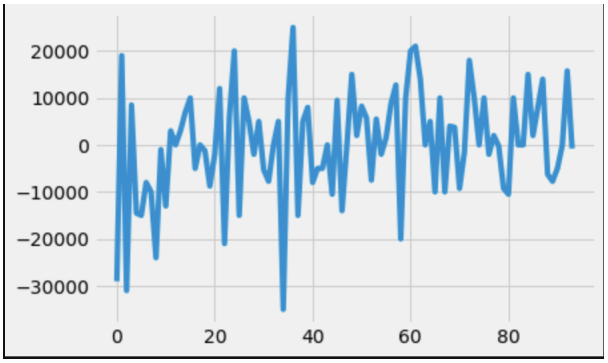
# Data Description:

The data will be split into two parts, the historical dataset, and the test dataset. The historical dataset has records from February 2008 - December 2015 that were provided by Zillow. We are provided with the monthly median sold price for housing in California, the monthly median mortgage rate, and the monthly unemployment rate. The test dataset will contain records from January 2016 to January 2016 of the same features. The specific dates for each monthly data will be written as the final day of the month. The historical dataset will be used for training, and the test dataset is to calculate the prediction accuracy for the final model.

# Exploration




The plot of the time series indicates that there is a linear trend in the second part of the series. The ADF test confirms that the series is not stationary as obvious from the plot. After differencing one time (as its monthly series, we are looking at month-over-month price changes), the ADF test indicates stationary time series. Having a stationary time series aids the modelling of the ARIMA family. A stationary time series implies the mean and variance are not functions of time. It also assumes that the correlation of two points is not a function of time but a function of the distance between them. We used the ADF test to check for stationarity. The plot after differencing once is given below.




The differenced time series seems to have a more pronounced seasonality though the variance is not constant throughout.

# Exponential Smoothing (ETS)

As shown in the exploration section, we saw non-stationary data. In fact we saw an overall decrease and then an increase in median prices as time went on. In addition from the

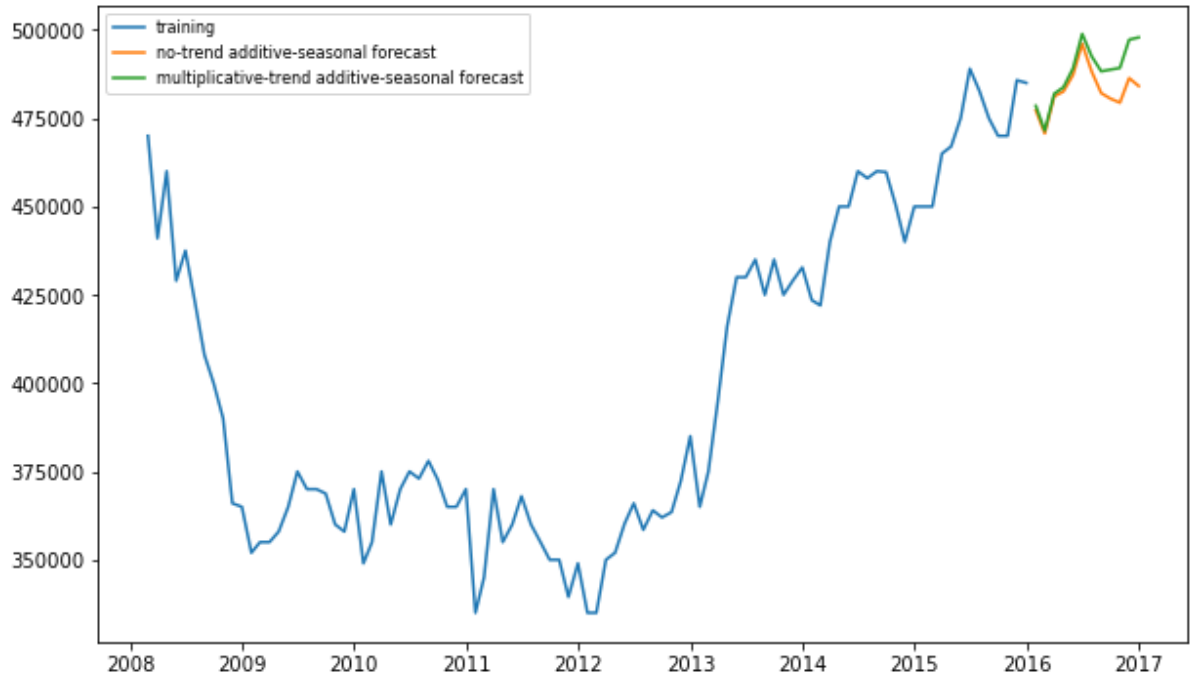
data description, the data is monthly that repeats on a yearly cycle. The cycle would suggest seasonality. Therefore, the data can be modeled with ETS using the triple exponential smoothing because there is a trend, a decrease and increase in overall price, and seasonality, a yearly cycle.

I will use two methods to select the ETS models, one is visually and the other is using the one-step cross validation. The decrease and increase suggests that the data could be quadratic or a higher order. The seasonality does not seem to be too different as shown in our one-step-differenced plot. So visually I would select a multiplicative trend parameter and an additive seasonal parameter. The one-step cross validation will use 80% of the historical data to start and forecast one step, each consecutive step will include hidden data in the 20%, until the entire hidden dataset has been forecasted. The prediction and the hidden data will then be compared to calculate the lowest Root Mean Squared Error (RMSE) and evaluate.

The results from the selection are the following:

| Trend Parameter | Seasonality Parameter | RMSE |
|---|---|---|
| Additive | Additive | 10999 |
| Additive | Multiplicative | 11151 |
| Multiplicative | Additive | 1.6464e+84 |
| Multiplicative | Multiplicative | 1.1834e+100 |
| None | Additive | 10339 |
| None | Multiplicative | 10759 |

The lowest validation RMSE score was the model with no trend and additive seasonality. Despite visually selecting multiplicative-additive as a model, the RMSE score did not suggest those parameters as optimal. Both models are shown below.




The no-trend model suggests that the upward trend since 2012 is not as important, while the multiplicative-trend model suggests that the upward trend would continue but at a decreased rate.

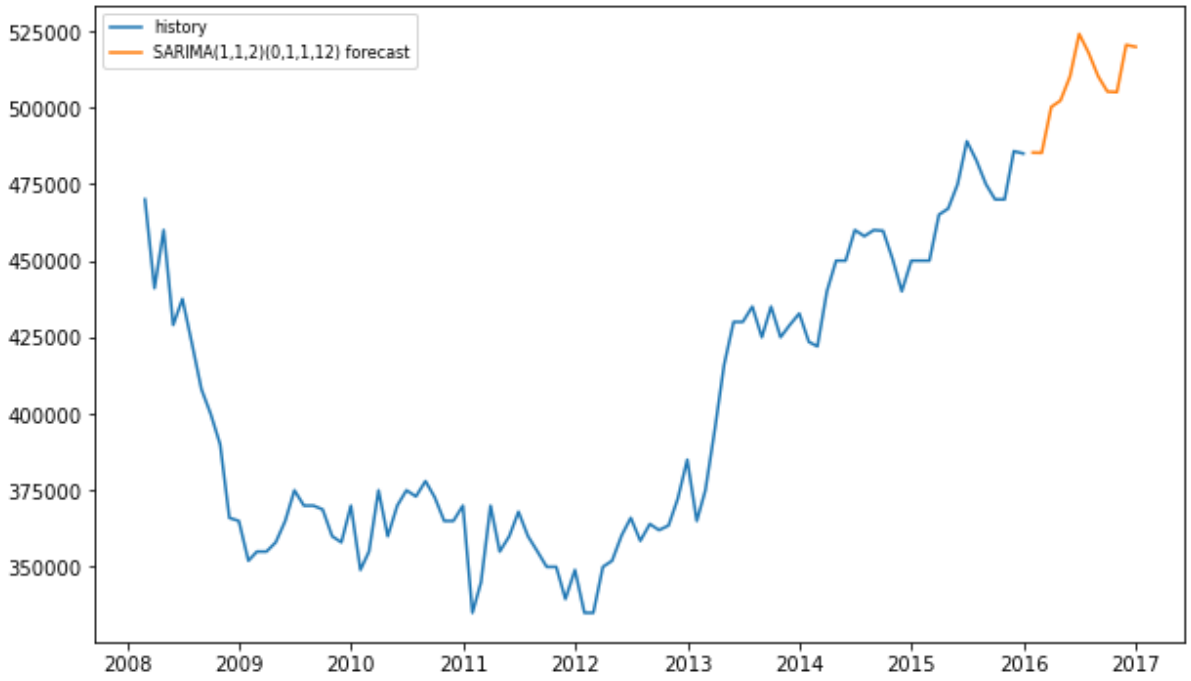
From this section we select two models, a no-trend additive-seasonality ETS model purely using the one-step cross validation RMSE, and a multiplicative-trend additive-seasonality ETS model visually from the time series plot.

# SARIMAX

Firstly, consider all the combinations of the possible exogenous variables, which are: mortgage rate, unemployment rate, and mortgage rate + unemployment rate. Then we perform parameter tuning and note the RMSE from one-step forward cross-validation for each potential exogenous variable set.

1. Taking unemployment rate as an exogenous variable, the best model we get is SARIMA(2,0,3)(2,2,0,12) with an RMSE of 12220.
2. Taking the mortgage rate as an exogenous variable, the best model we get is SARIMA(1,1,2)(0,1,1,12) with an RMSE of 9753.
3. Taking both unemployment rate and mortgage rate as exogenous variables, the best model we get is SARIMA(0,1,0)(0,1,0,12) with an RMSE of 10038.

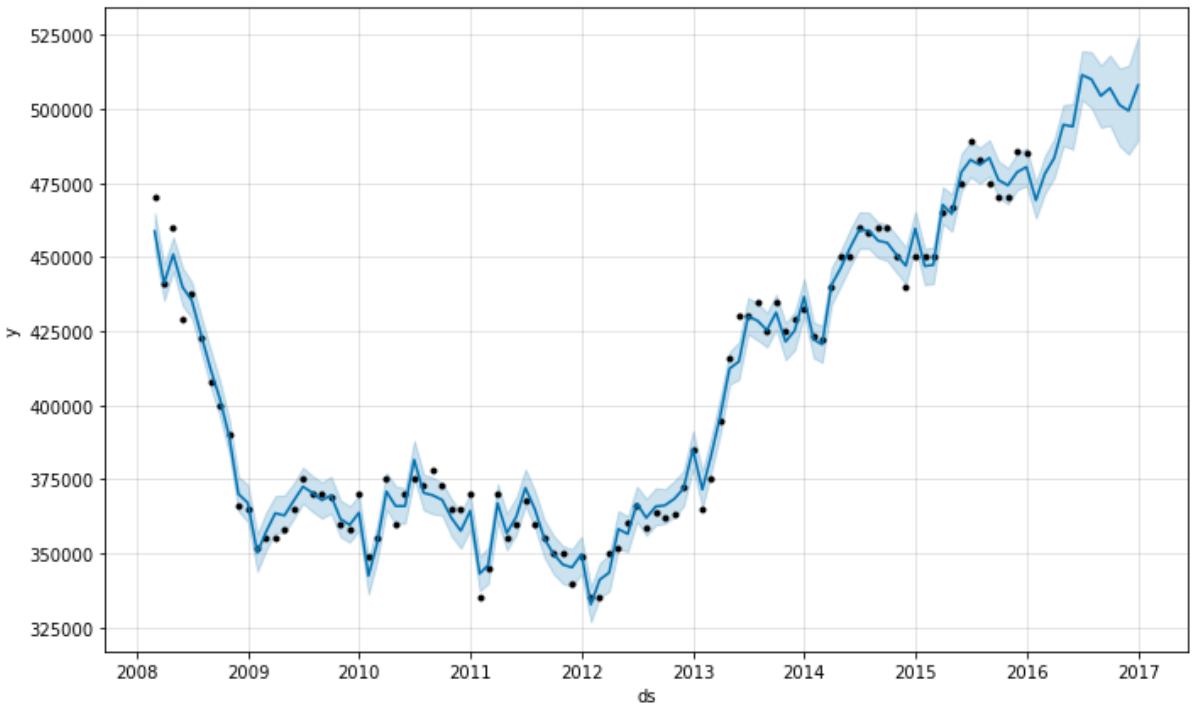
According to the collected RMSE, the best model overall is SARIMA(1,1,2)(0,1,1,12), use this model to forecast the median sell price in 2016, we have below plot:




# FB Prophet

We consider the univariate case of modelling the median home prices as Prophet's performance with multivariate time series is not optimal. With the Prophet model, we don't need to difference the time series as the model works out the change points.

Fitting a Prophet model is ideal with data recorded daily but we have monthly observations. We have made the appropriate adjustments with seasonality set to 12 months (one entire year). The fitted model along with the forecasts is below.




The components of the forecast by Prophet are:


RMSE from the fitted Prophet model is 8,603.

# Comparison:

Test conducted on forecasted values for Jan-Dec 2016.

| Univariate/Multivariate | Model | Test-RMSE |
|---|---|---|
| Univariate | no-trend, additive-seasonality ETS | 21,605 |
| Univariate | multiplicative-trend, additive-seasonality ETS | 15,860 |
| Multivariate | SARIMAX(1,1,2),(0,1,1,12) | 12,578 |

| | | |
|---|---|---|
| | (Median Price + Mortgage Rate) | |
| Univariate | Prophet | 8,603 |

# Conclusion:

We used three different methods to forecast the Median Sale Price of California houses in 2016 using historical data from February 2008 - December 2015. Based on the plots and RMSE scores on the test set, we conclude that the best model is the univariate Prophet model with a Test RMSE of 8,603. The model was the most accurate in forecasting 2016 median sale prices based on the historical data provided.