Zillow Housing Price Prediction

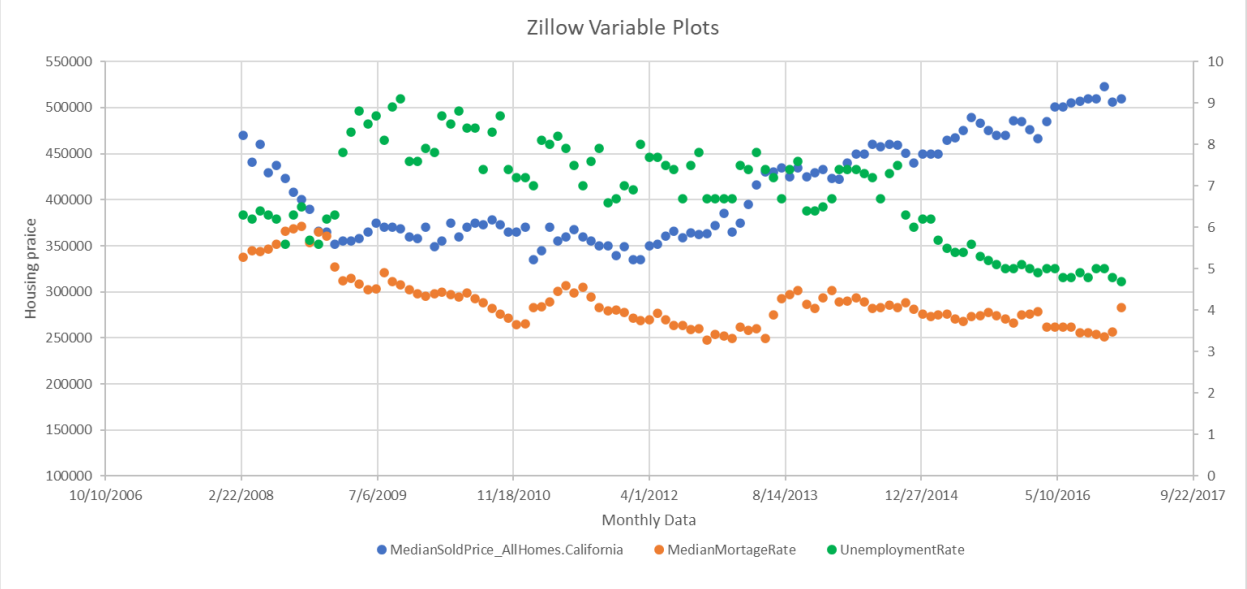
Jan -Dec 2016

Kris Knapp, Meilin Li, and Jeffery Ott

Introduction:

Using a combination of monthly historical data of California median housing prices, median mortgage rate, and unemployment rate from Feb 2008 to Dec 2015, we attempted to predict the future housing market for 2016. We utilized three different time series models (FBPROPHET, ETS, and SARIMAX) to calculate this data. Our approach was to exhaustively try all possible variable combinations presented to use and choose the best of the models selected from RMSE cross-validation score.

The Data



Target:

Median Sold Price All Homes California

Additional Variables:

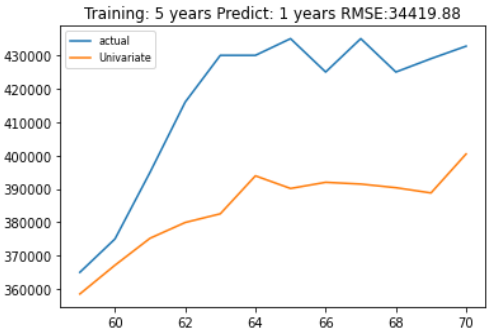
Median Mortgage Rate

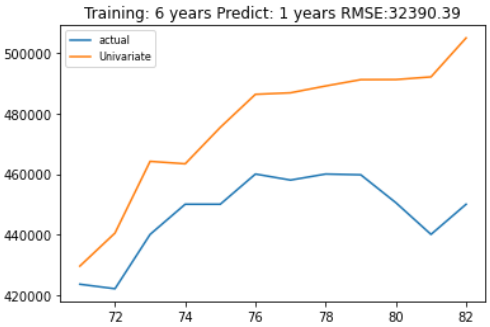
Unemployment Rate

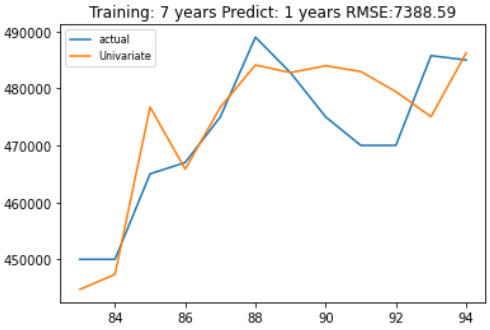
FB Time series model testing

In this section, we attempted to fit the data using the FB prophet model, a time series forecasting procedure made by Facebook. In our implementation, we tried both the Univariate models and Multivariable monthly Prophet models in our implementation. To see the effectiveness of these models, we cross-validated them below by training for 5 years, then predicting a window of 1 year and rolling that year. For cross-validation, it seemed as if each model drastically improved with the last two testing years for both Univariate and Multivariate Models

Univariate Cross Validation

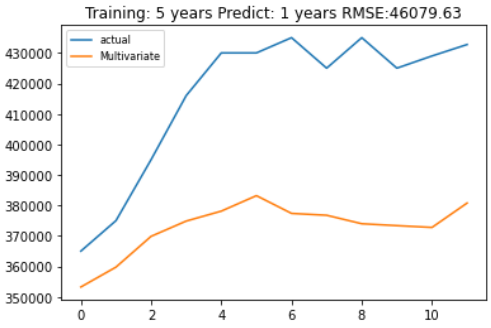


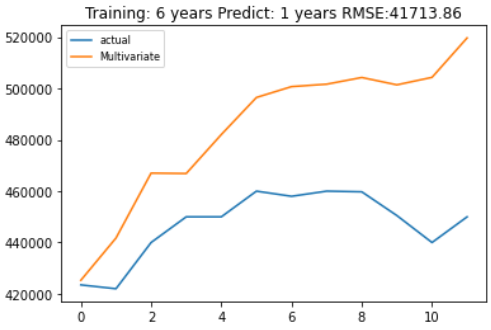


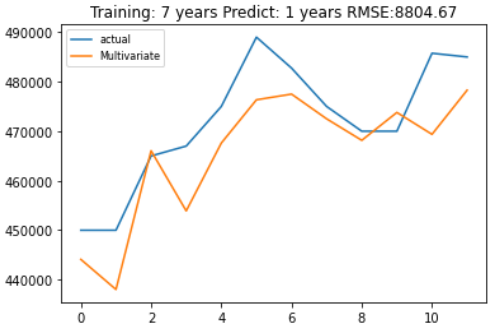


Mean RMSE: $24,732

Multivariate Cross Validation







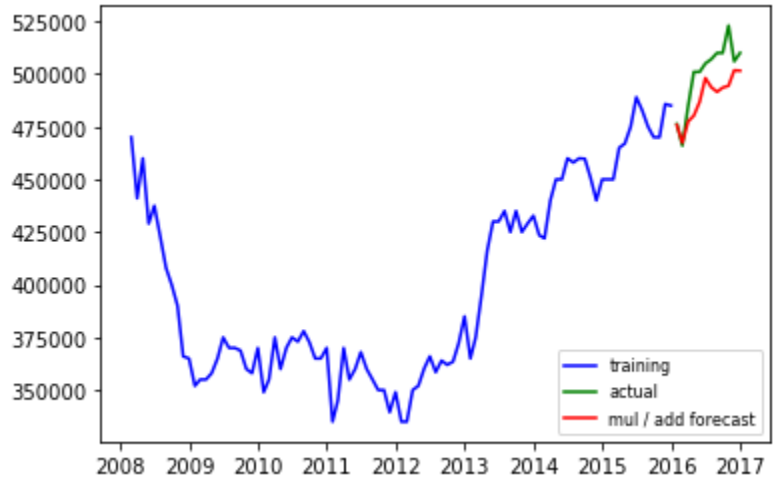
Mean RMSE: $32,199

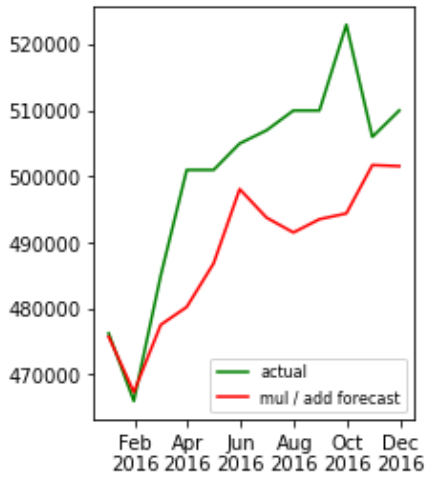
ETS Time series model testing

Next, we attempted to fit the data using Holt-Winters triple exponential smoothing models. We compared four different models, each fit with the trend and seasonality parameters set to all possible combinations of 'additive' and 'multiplicative'. Frequency was kept at the default value of monthly, with seasonal periods set to 12 months and the damped trend parameter left on True for all models.

frequency	seasonal period	damped trend	trend	seasonality	model avg c-v RMSE
Monthly	12 months	True	**ADD**itive	**ADD**itive	46915
Monthly	12 months	True	**ADD**itive	**MUL**tiplicative	40154
Monthly	12 months	True	**MUL**tiplicative	**ADD**itive	**29876**
Monthly	12 months	True	**MUL**tiplicative	**MUL**tiplicative	32672

The best RMSE value of $29,876 was achieved during cross-validation by the ETS model using the 'multiplicative' trend and 'additive' seasonality. After re-fitting an ETS model with the same parameters on the entire training dataset and using it to make forecasts into the 12-month period of the test data (2016), the results had an RMSE score of $11,205 on the test period.





SARIMAX Time series model testing

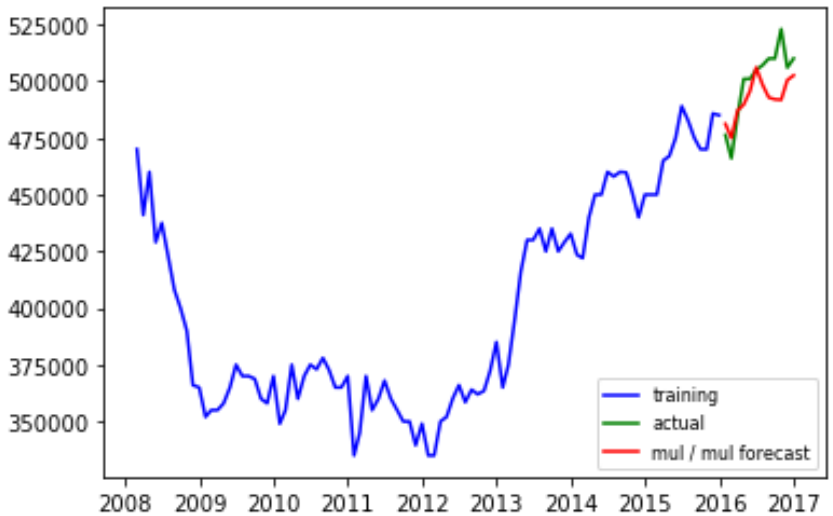
Finally, we fitted the Autoregressive Integrated Moving Average Model (ARIMA), the most widely used forecasting method in timer series data forecasting. From the time series plot we generated below, we can see that there is an obvious seasonality existing in the data (one year as a period), which is also corresponding to reality.

After ADF test and seasonal differencing, we decided on the trend parameter d = 1 and seasonal parameter D = 1. Next, we performed the grid search to obtain other parameters. The two candidates we got from the SARIMA family are: SARIMA(1,1,0)(0,1,0,12) for univariate model and SARIMAX(1,0,0)(0,1,3,12) for multivariate model.

Then we performed both one-step cross-validation and twelve-step cross-validation to see the cross-validation RMSE score and decide the best one in the SARIMA family. And the scores are:

	1 step	12 step
SARIMA(1,1,0)(0,1,0,12)	10032	20607
SARIMAX(1,0,0)(0,1,3,12)	25242	37299

And the prediction result of the SARIMA is as the graph shows below.

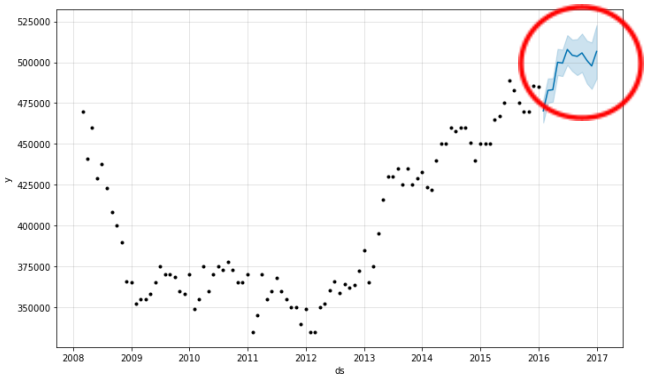


Test score(RMSE) is: 12944

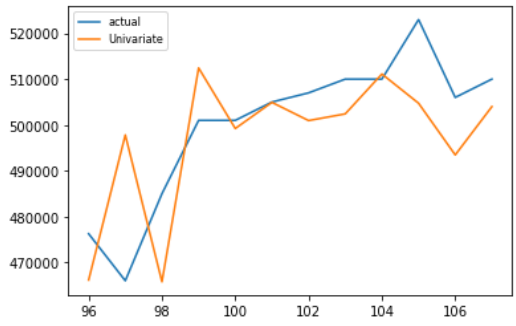
Final Model Selection-Mulivariate FBProphet

Univariate Model Forecast

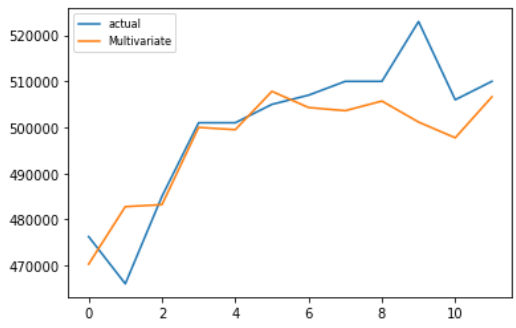
Multivariate Test Forecast



With the similar cross-validation error scores and the significant drop in RMSE with both models, it was a simple procedure to run both Univariate and Multivariate FBprophet on the test data in order to determine the final model. We found that the **Multivariate FB prophet utilizing both the Unemployment and MedianMortageRate within the mode**l yielded the best RMSE results. From this model, it seems that we are able to guess the median house price with only a small margin of error



This univariate fit was a little noisy at the beginning but overall ended with an **RMSE of $13,663.** This means that our average error prediction on each of the houses is around $13663 which is a pretty good estimate when compared to the average housing price.



The Multivariate fit was more on point, but I need to make sure I had the additional variables lined up with the correct dates. With this method, the **RMSE was $7720**