# Time Series Final Report

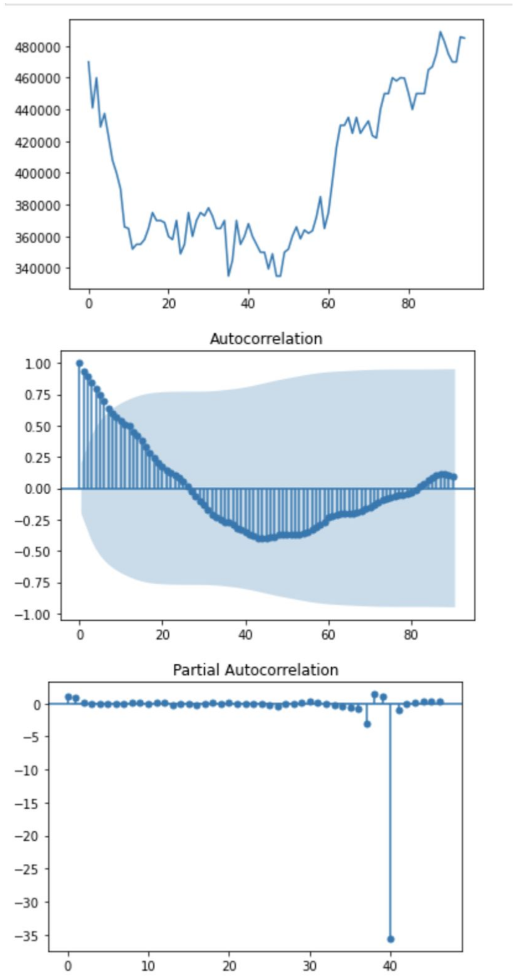
Catie Cronister, Yueling Wu, Max Shinnerl

## Introduction

In this project, we were tasked with predicting median home prices using time series techniques covered in our time series lectures. This complex time series of housing prices could either be modeled using univariate methods or with multivariate methods as we were given two exogenous variables -- unemployment rate and median mortgage payment. Each person in our group fit at least one algorithm and tuned the model before trying the model on the validation set. Once we had our final model for each algorithm, we tested it on the test data.

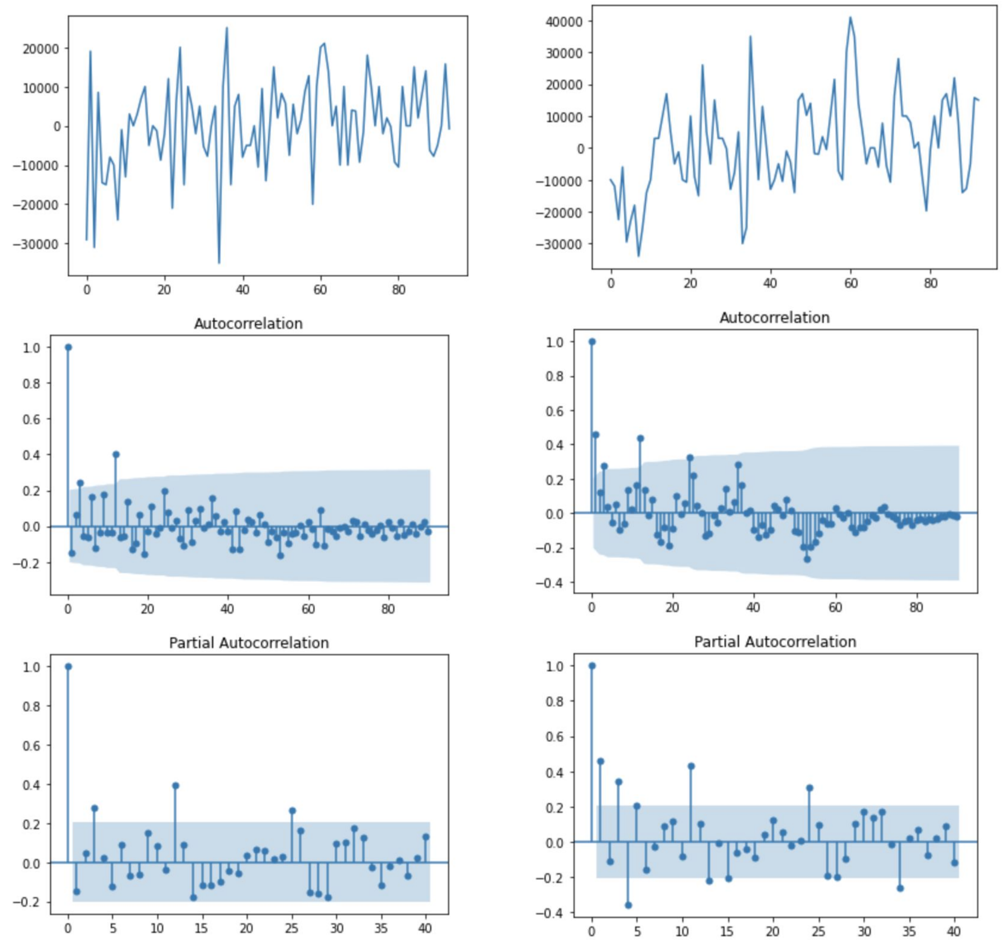
## SARIMA

We first fit a SARIMA model we could capture the trend, the seasonal component and the random noise of our data. An initial plot of Median Sold Price versus time showed us there was evidence of a trend and possible seasonality. Because of this we wanted to pick a model type that allows for detrending. Our initial thoughts regarding this data was that it follows a second order trend so we should difference once. We can also see that it is most likely an order two trend because of the shape of the ACF plot.



Differencing once produced the following graphs and an ADF score of -3.088139 and a p-value of 0.027443. So this is now a stationary time series, however I was interested in the second order differencing -- but we did not want to overfit the model. However we decided to also look into differencing twice. This produced an  ADF score of -3.374527 and a p-value of 0.011859.

We then considered the ACF and PACF plot to determine a possible seasonal trend value, m. This seemed to be about 12 or 13.

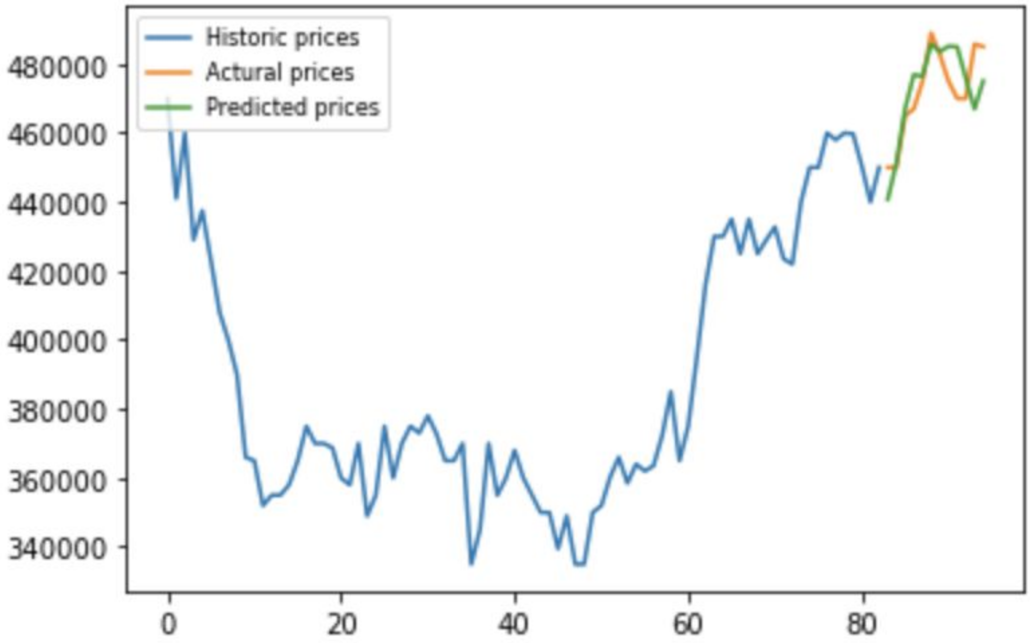


d=1

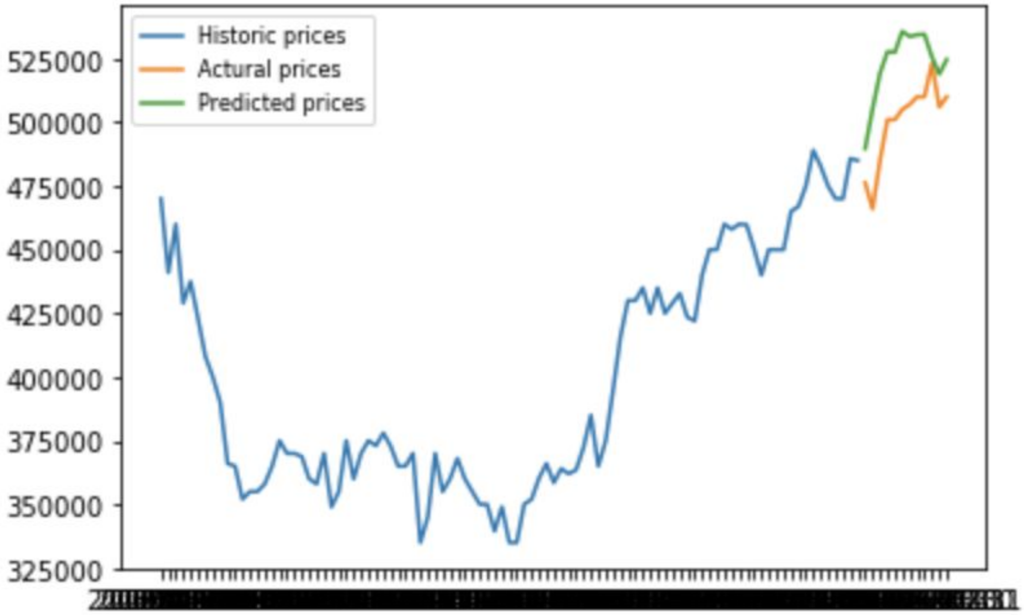
d=2

From here we decided to include d=1 and d=2 in our grid search to minimize BIC for the SARIMA model. Running BIC SARIMA cross validation with the parameters m_values=[12, 13], d_values=[1,2], p_values=[0,1,2,3], q_values=[0,1,2,3], Q_values=[0,1,2,3], P_values=[0,1,2,3] and D=1 gave us a best model and BIC of:
(0, 2, 1), (0, 1, 2, 12) and 16.029332740929885

We then fit the model and use all the data before 2015 as history and data after that as the validation set. We forecasted one step at a time and produced the following graph:



This model gives us an RMSE of 9192.337995183163, which considering the scale of the prices, is not too shabby. The model was not quite as effective on the test data as it produced the following graph with an RMSE of 10580.236433484737 but still relatively successful.



Tabular Forecasting Results for SARIMA:

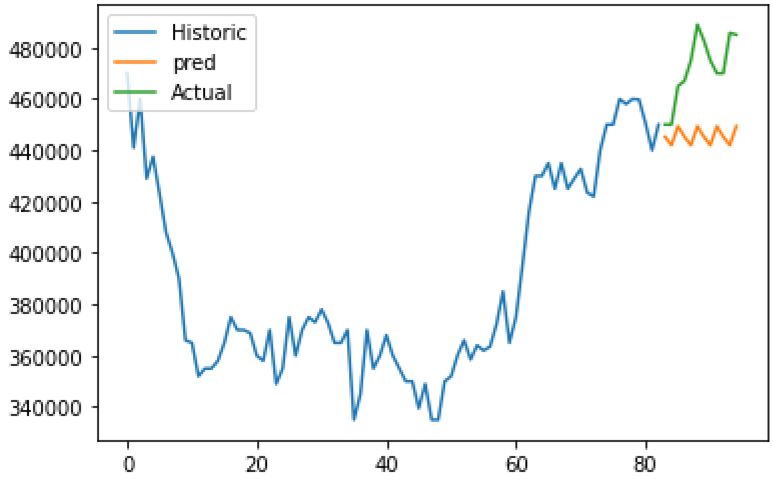
| Month | pred_price |
|---|---|
| **2016-01-31** | 466849.0707497960 |
| **2016-02-29** | 473504.4390554480 |
| **2016-03-31** | 488687.91397339500 |
| **2016-04-30** | 497781.12132832200 |
| **2016-05-31** | 498103.80117814800 |
| **2016-06-30** | 506727.62285251700 |
| **2016-07-31** | 504897.6423975650 |
| **2016-08-31** | 505733.34440762800 |
| **2016-09-30** | 506058.23134750700 |
| **2016-10-31** | 497130.56720764500 |
| **2016-11-30** | 489481.07927028500 |
| **2016-12-31** | 496472.09812174800 |

## ETS

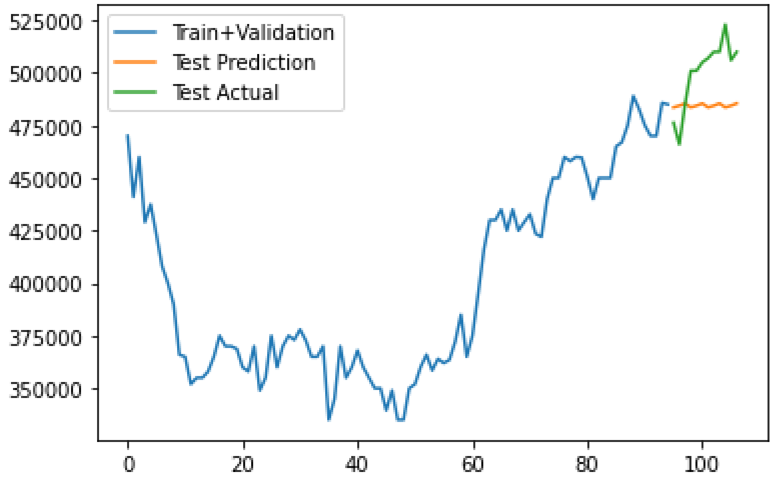
Although the SARIMA model was pretty good, to be thorough we fit an ETS and a Prophet model as well. We decided to include an ETS model as we wanted to see if a more general model was able to achieve similar results as the more specific SARIMA. For the ETS model, we did a form of grid search.  With m ranging from 2 to 14, trend being either None, ‘additive’, or ‘multiplicative’, and seasonal being ‘additive’ or ‘multiplicative’, we tried all different permutations and (by MAPE score) landed on the following as the best ETS model:

Trend=None, Seasonal="multiplicative",m=3,damped=False

Against the validation set, this scored a MAPE score of 0.0555.



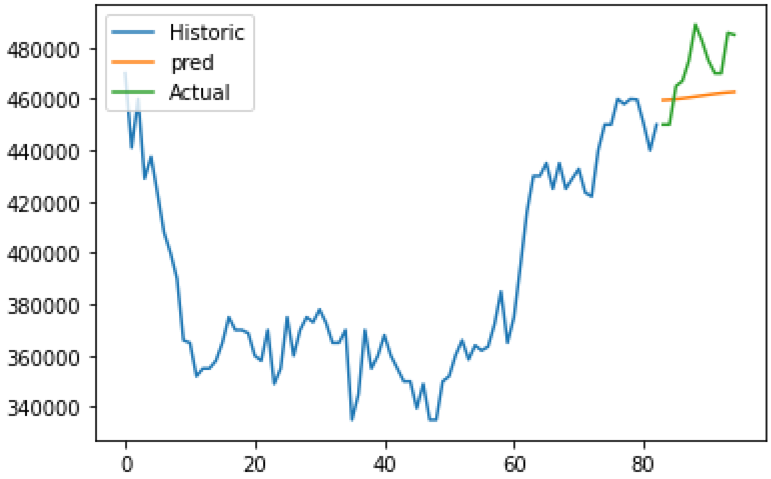
Against the test set, this scored an RMSE of 21947, significantly worse than the SARIMA result.



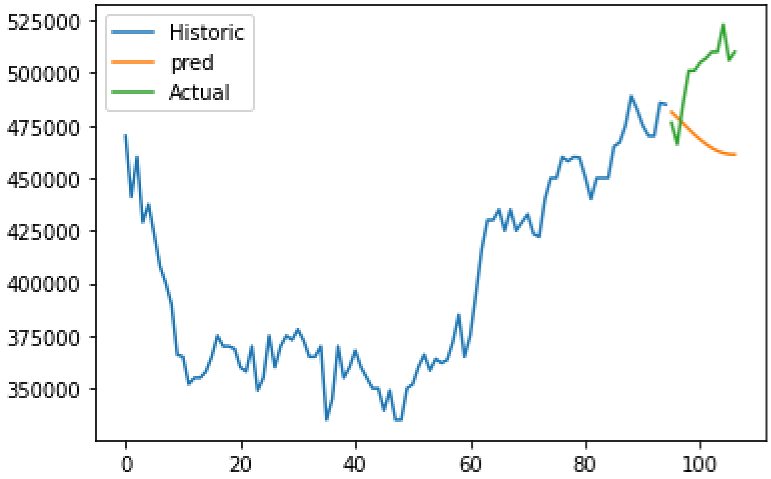
## Prophet

Since Prophet tends to perform better for long-term forecasting, it seemed unlikely that it would perform well in this scenario. Similarly, Prophet works best on daily data and this data was collected monthly, so we did not have high hopes for the model. However, just to be thorough, we also fit a Prophet model with the default hyperparameters.  As expected, it did not do well.

Against the validation set, it got a MAPE score of 3692.

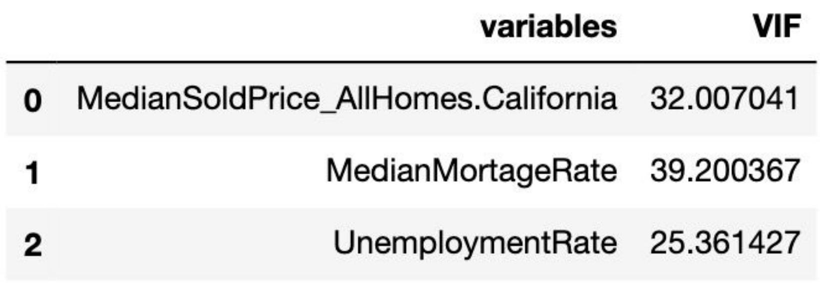


Against the test set, it got an RMSE of roughly 38070.



## SARIMAX

We verify the multicollinearity among Median Sold Price, Median Mortgage Rate and Unemployment Rate by inspecting VIF. Also it is reasonable to think that mortgage rate and unemployment rate will affect housing prices. So we take mortgage rate and unemployment rate as exogenous variables, applying SARIMAX to fit the price data.

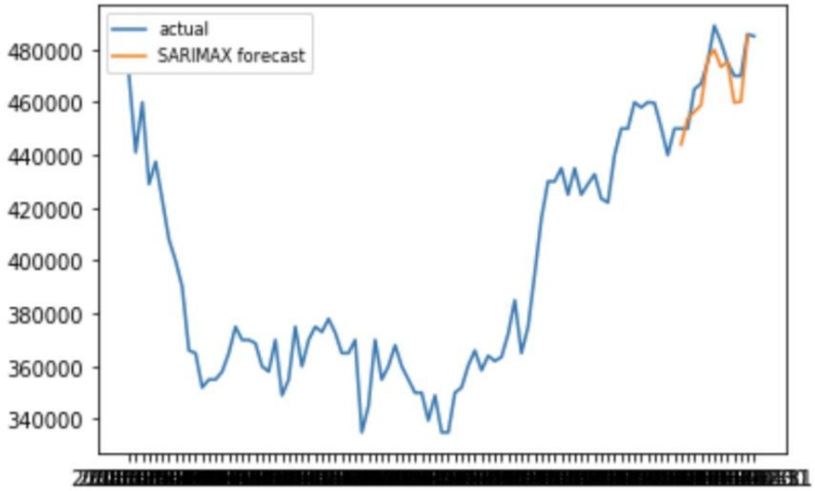


| | variables | VIF |
|---|---|---|
| 0 | MedianSoldPrice_AllHomes.California | 32.007041 |
| 1 | MedianMortageRate | 39.200367 |
| 2 | UnemploymentRate | 25.361427 |

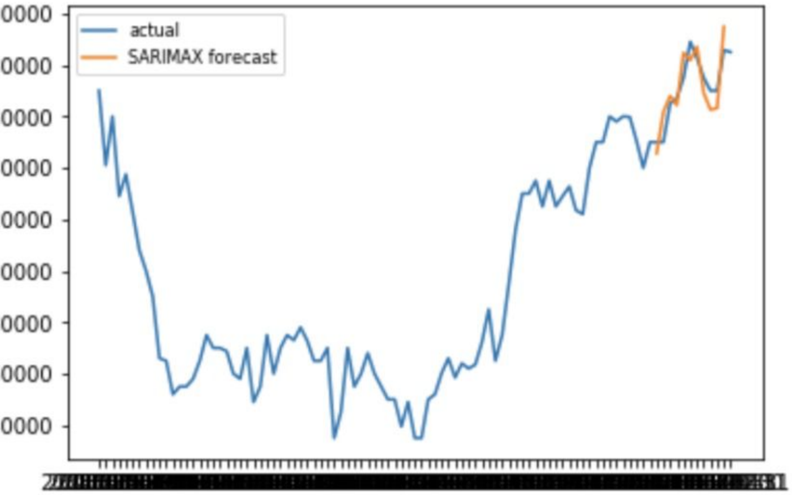
For better prediction, we first normalize the features and take the house price of next month as future price.

For consistency, we tried both m=3 and m=12 here. Since the housing price is recorded monthly, we assume there exists a yearly trend, though the time series plot, acf plot and pacf plot cannot show the seasonality clearly. When testing ETS, the best model chooses m=3, so we are also going to test it in SARIMAX.

When m=3, the best SARIMAX model is SARIMAX(1,0,0)(0,1,1)[3]. The RMSE score on the test set is 7149.654 and the MAPE is 0.012965.



When m=12, the best SARIMAX model is SARIMAX(0,0,2)(1,1,2)[12]. The RMSE score on the test set is 7075.799 and the MAPE is 0.013868 .



SARIMAX(0,0,2)(1,1,2)[12] has a better RSME score while SARIMAX(1,0,0)(0,1,1)[3] has a better MAPE score.

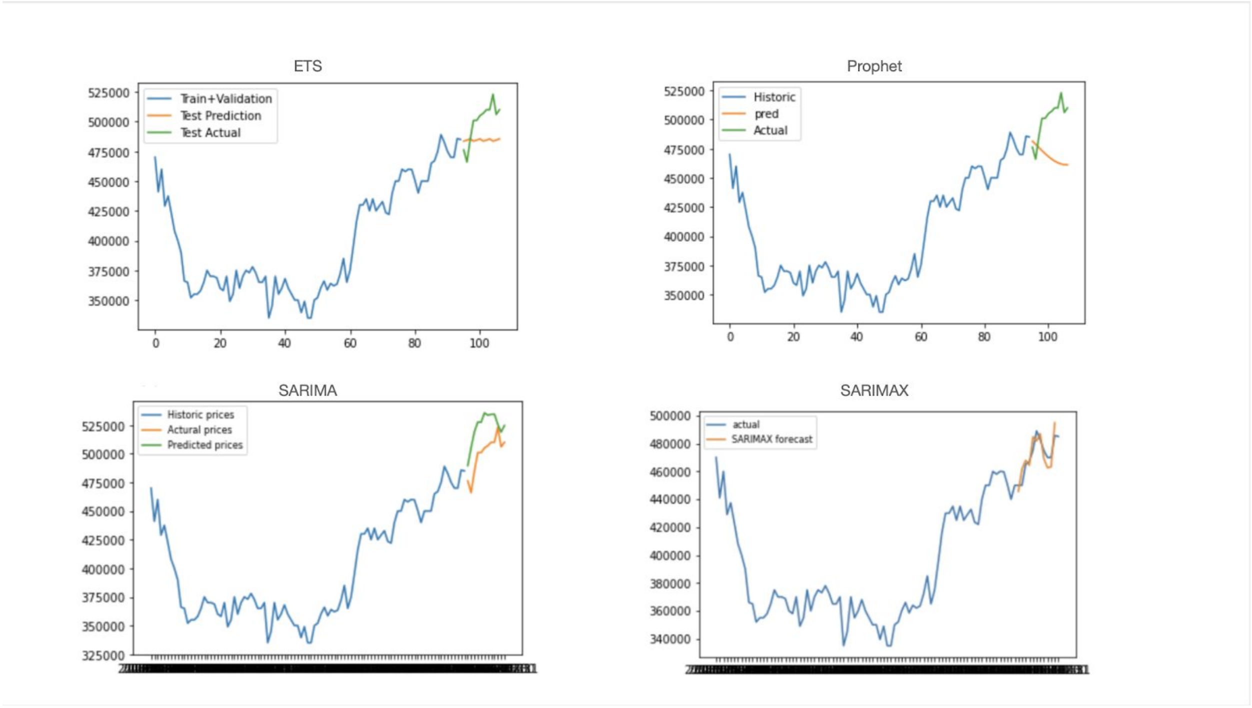
## CONCLUSION

| MODEL | **SARIMA**<br>(0, 2, 1), (0, 1, 2, 12) | **ETS** | **Prophet** | **SARIMAX**<br>(1,0,0)(0,1,1)[3] | **SARIMAX**<br>(0,0,2)(1,1,2)[12] |
|---|---|---|---|---|---|
| **RSME** | 10580.236 | 21947 | 38070 | 7149.654 | 7075.799 |
| **MAPE** | 0.01566 | 0.0555 | 3692 | 0.012965 | 0.013868 |

Throughout the course of this project, we fit four different time series  algorithms -- SARIMA, ETS, SARIMAX and Prophet. We then “tuned” these models through order selection or hyperparameter tuning. After getting a best model for each algorithm, we tested the model on the test set. From this process, we found that SARIMAX had the overall lowest RMSE and MAPE and SARIMA had the lowest RMSE and MAPE among the univariate algorithms.

For our final model we chose SARIMAX(0,0,2)(1,1,2)[12] as, again, we saw it had the lowest RMSE and MAPE. Thus, we can say that median home price is easier to model when you include the exogenous variables median mortgage payment and unemployment rate. This was a great example of the interconnectedness of the housing market and that many factors go into the pricing of homes. Because of this interconnectedness, univariate algorithms were not strong enough to model this time series, but the multivariate SARIMAX was able to model this complex time series.

### Final Model for Each Algorithm and its Performance on the Test Set


ETS
Train+Validation
Test Prediction
Test Actual
Prophet
Historic
pred
Actual
SARIMA
Historic prices
Actural prices
Predicted prices
SARIMAX
actual
SARIMAX forecast