Final Project

Monthly Median Sold Price for Housing in California

Stephen Embry, Matthew Hui, Shiqi Tao

Description of dataset:

History data from Feb 2008 - Dec 2015 were separated into train and validation dataset, where train data is from 2008-02-29 to 2014-12-31, and validation data is from 2015-01-30 to 2015-12-31.

In order to forecast median sold housing price in this dataset, we tried two general approaches: 1) univariate time series; 2) multivariate time series.

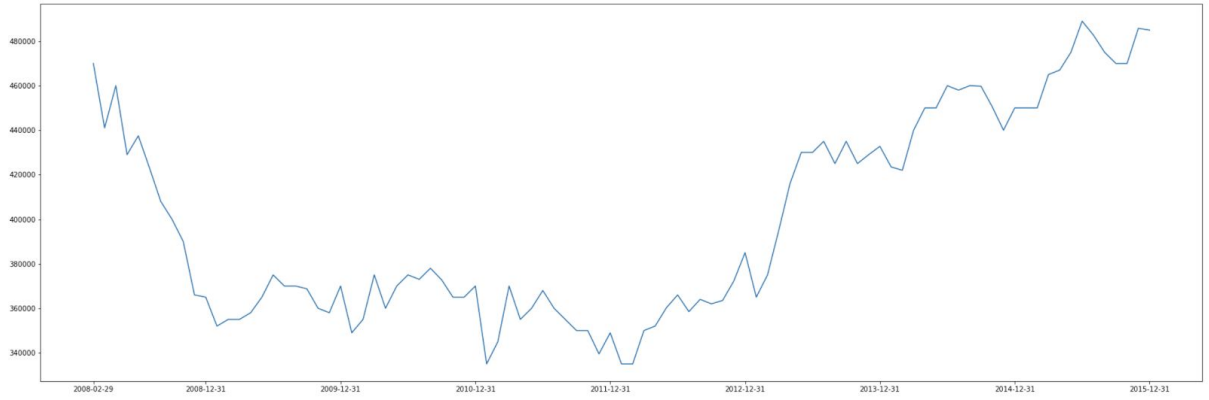
Because we wanted to first have a general picture about how median sold housing price alone would perform on forecasting near future, and then add other two possible factors that could affect housing price.

➢ Univariate

1. SARIMA:

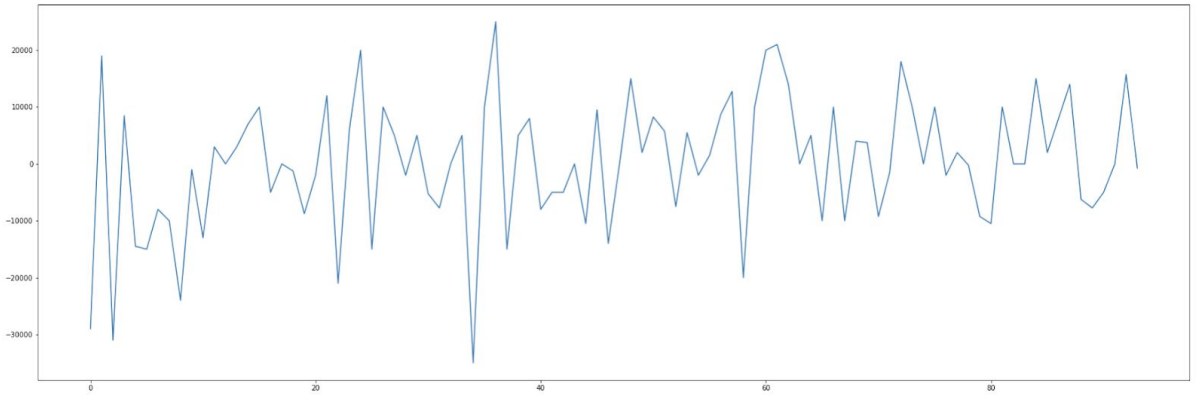
With time series plot on median sold housing price, acf plot and pacf plot, we first observe an obvious curve trend (decreasing first and then increasing), and a yearly seasonality with higher price during each summer.

Time series plot:



After one time differencing the data, we already got stationary data and p-value from the ADF test verified it (p-value = 0.027).

Detrend once:

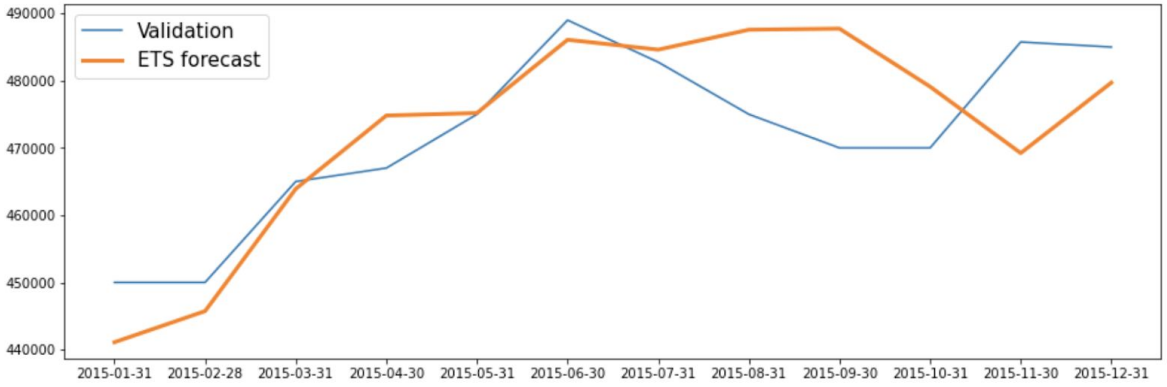


Therefore we decided to directly grid search the best SARIMA model, with search range:

p_value	[0, 1, 2, 3, 4]
d_value	[1, 2]
q_value	[0, 1, 2, 3, 4]
P_value	[0, 1]
Q_value	[0, 1, 2, 3]
m	12
D	1

Grid search gave the first candidate model SARIMA (2, 1, 4) (1, 1, 1, 12), with RMSE = 9257 on validation dataset. Here is a forecasting plot to show performance of this model. We can observe that the SARIMA model is capturing the near future's pattern better than long term (forecasted on 2015 data).

Candidate 1: SARIMA forecasting



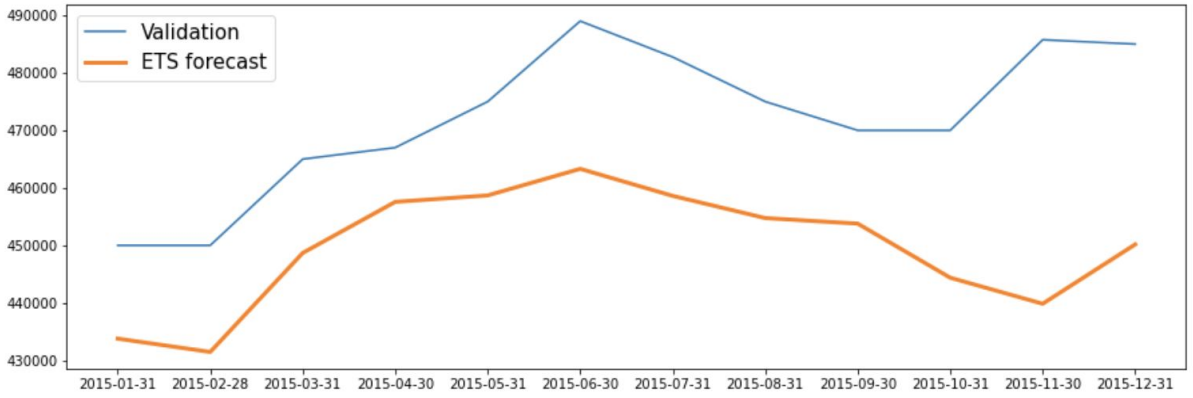
2. ETS

After having our first SARIMA model, we tried to fit a ETS model to check whether the RMSE can be improved.
For the ETS family, we used cross validation to compare rmse value of all possible combinations of trend and seasonality with m=12.

MODELS	TREND	SEASONAL	m	DAMPED	RMSE
model_1	additive	additive	12	True	10079
model_2	multiplicative	additive	12	True	Nan
model_3	additive	multiplicative	12	True	10033
model_4	multiplicative	multiplicative	12	True	Nan
model_5	None	additive	12	False	9968
model_6	None	multiplicative	12	False	10223

Obviously model_5 with no trend and additive seasonality has the lowest rmse value. So we chose this model to do a forecast and compare it with the validation set. The final rmse value is 24360, which is much higher than the SARIMA model.

Candidate 2: ETS forecasting (Trend: None, Seasonal: additive)



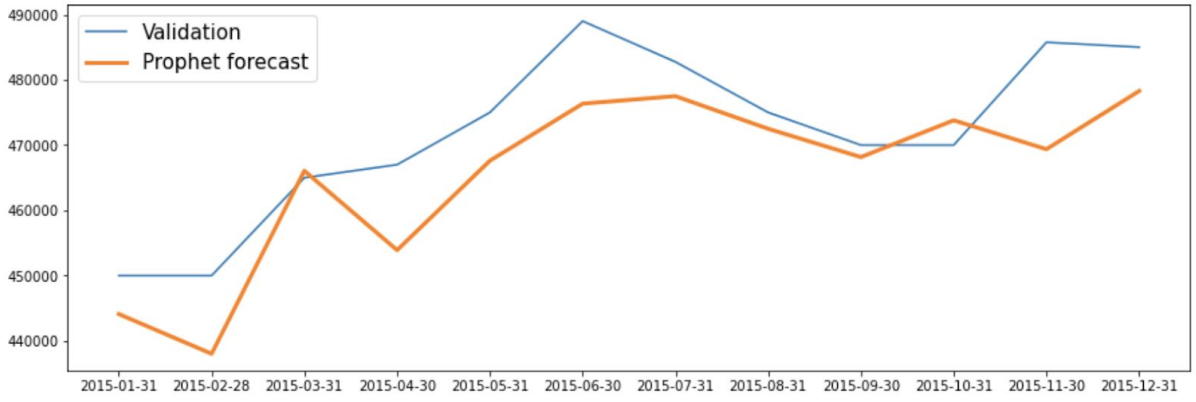
The forecasting plot shows that the prediction of our ETS model is always below the true values, however, it captures the pattern well.

3. Prophet (adding regressors)

We fit a Prophet candidate model adding two regressors- median mortgage rate and unemployment rate. This candidate model is performing as a multivariate time series rather than univariate.

With the Prophet model we also tried to add yearly seasonality and holidays, but the rmse increased on the validation data. As a result, we only added two regressors that could affect median sold housing price, and the final RMSE value comparing prediction to validation is 8805.

Candidate 3: Prophet forecasting

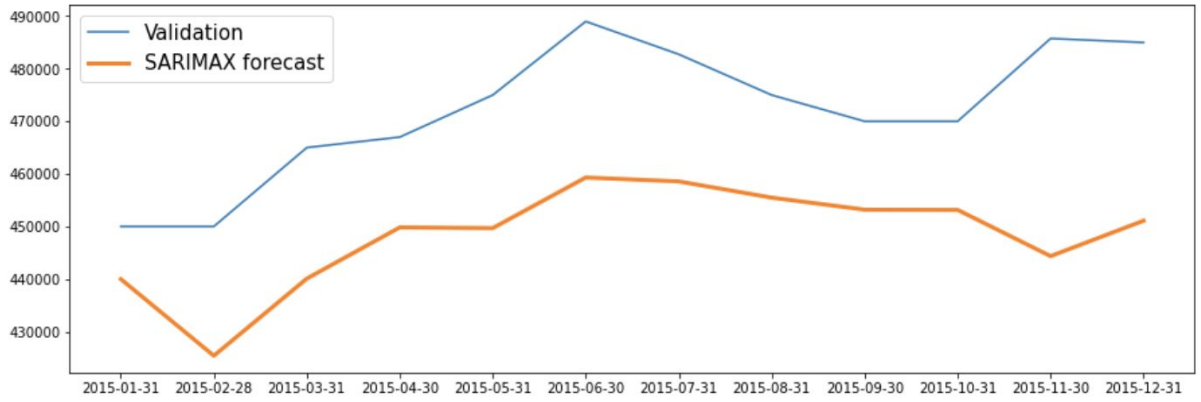


➢ Multivariate

1. SARIMAX

We wanted to add both the Median Mortgage Rate and Unemployment rate of the month to a SARIMA model. We did a grid search using the same parameters as the SARIMA model above in order to come up with a candidate model. When using our grid search we decided to find the lowest BIC in hopes of finding a simpler model that still can perform well. Grid search gave us the candidate model SARIMAX (2, 1, 2) (0, 1, 0, 12). We get a validation error of 25039 using this model.

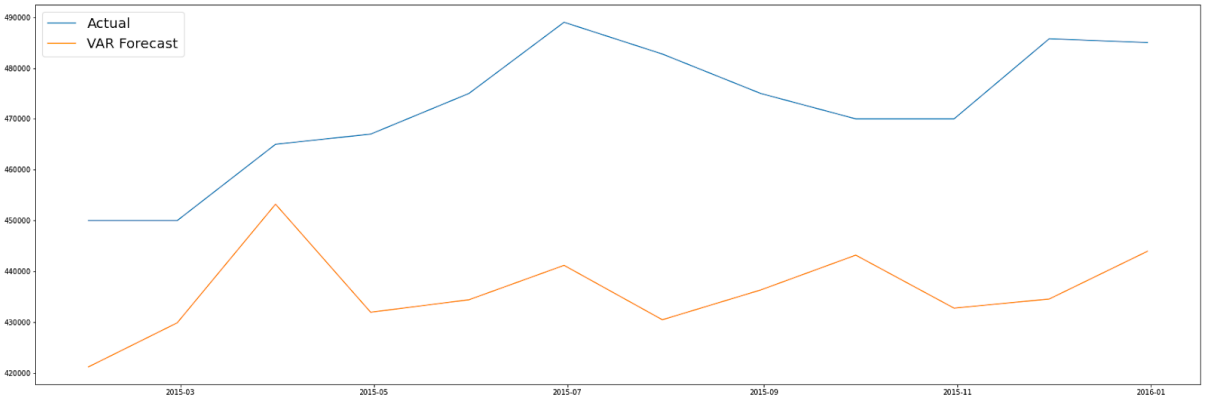
Candidate 4: SARIMAX forecasting



2. VAR

When fitting our VAR model, we once again used both mortgage and unemployment rate in the model. When fitting this model, we used a second differencing and looked at BIC for order selection. With this model, we got a validation RMSE of 37816.

Candidate 5: VAR forecasting



Summary of Findings:

Below we provided the validation error for our 5 candidate models based off of different algorithms.

Model	RMSE
SARIMA	9257
ETS	24360
Prophet	8805
SARIMAX	25039
VAR	37816

Final Model:

For our final model, we decided to use the Prophet model while using mortgage rate and unemployment rate as regressors. We chose the Prophet model because it was the model with the lowest validation error. When applying the Prophet model to our test dataset, we got a final test RMSE of 7720.

Time Series plot with Forecasted Test:

Date	Predicted	Actual
1/2016	467817	476250
2/2016	477860	466000
3/2016	482594	485000
4/2016	499735	501000
5/2016	498958	501000
6/2016	508035	505000
7/2016	508590	507000
8/2016	505519	510000
9/2016	507532	510000
10/2016	503400	523000
11/2016	499605	506000
12/2016	504624	510000