Time Series Group 8 Project

Sihan Chen, Jingxian Li, Xu Liu, Kathy Yi

1. Description of the problem

Goal: To predict the median sold price from January 2016 to August 2017
The data contains the following monthly information from 2004 to 2017 of California:

1. median sold price
2. median mortgage rate
3. unemployment rate
4. median rental price

We have 72 missing values for median rental price at the beginning of the data, which is from the year 2004 to 2009. We mainly use three ways to deal with missing values. Also, we want to explore different types of models and compare them, using the same feature set.

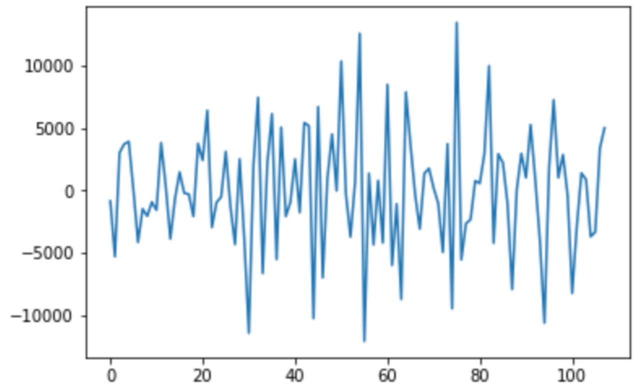
2. Description of the methods

Looking at the distributions of all variables, we found them to be normal and non-skewed; no feature transformations were needed.

Since we have missing values in our data, we first dropped all the missing values of median rental price since they were originally successive. Afterwards, we split the data at 0.85-0.15 train-test principle and implemented the following methods:

1. SARIMA
   After second-order differencing with, the data looks stational and we took a further differencing with lag=12 because it's a monthly data.

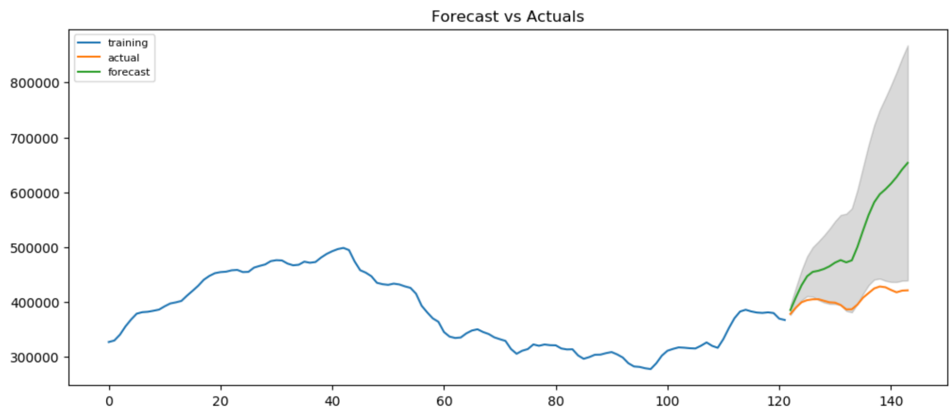
Then we fit the data with SARIMA model and got the following result:




```
Results of Dickey-Fuller Test:
Test Statistic                  -4.447752
p-value                          0.000244
#Lags Used                      13.000000
Number of Observations Used     94.000000
Critical Value (1%)             -3.501912
Critical Value (5%)             -2.892815
Critical Value (10%)            -2.583454
dtype: float64
```

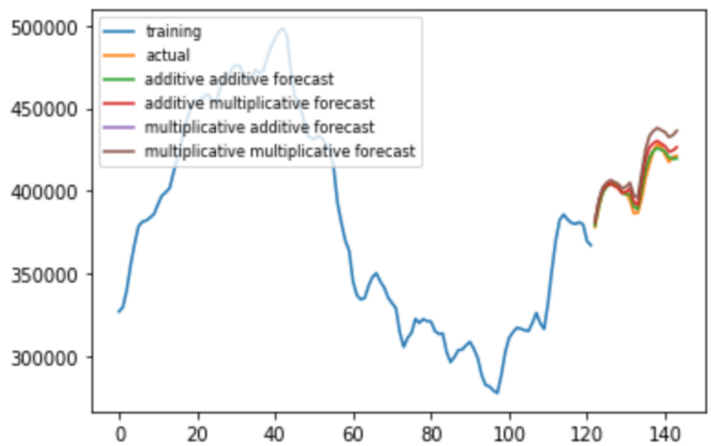
```
                                 Statespace Model Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  122
Model:             SARIMAX(1, 2, 1)x(0, 1, 0, 12)   Log Likelihood               -1061.573
Date:                            Sun, 08 Dec 2019   AIC                           2131.146
Time:                                    14:29:02   BIC                           2141.874
Sample:                                         0   HQIC                          2135.496
                                            - 122
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     32.2938     15.890      2.032      0.042       1.150      63.438
ar.L1          0.7155      0.069     10.336      0.000       0.580       0.851
ma.L1         -0.9988      0.126     -7.904      0.000      -1.247      -0.751
sigma2      2.144e+07   5.39e-06   3.97e+12      0.000    2.14e+07    2.14e+07
===================================================================================
Ljung-Box (Q):                       49.87   Jarque-Bera (JB):                 1.86
Prob(Q):                              0.14   Prob(JB):                         0.40
Heteroskedasticity (H):               1.55   Skew:                             0.01
Prob(H) (two-sided):                  0.19   Kurtosis:                         3.64
===================================================================================
```

This model passed the model diagnosis and the validation RMSE is 127,775.68, here is the validation plot:

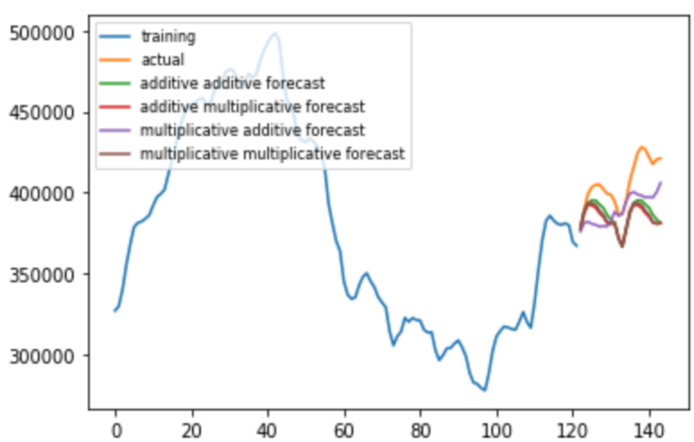


2. TES
   Four possible combinations of trend and seasonality were tried with TES.

   This is when we don't damp the trend:

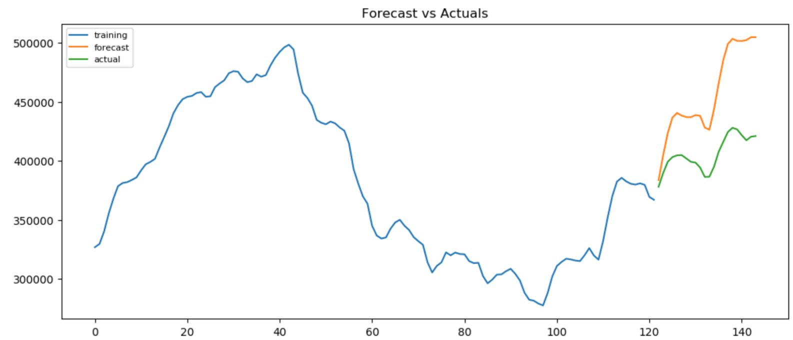


   This is when we damp the trend:

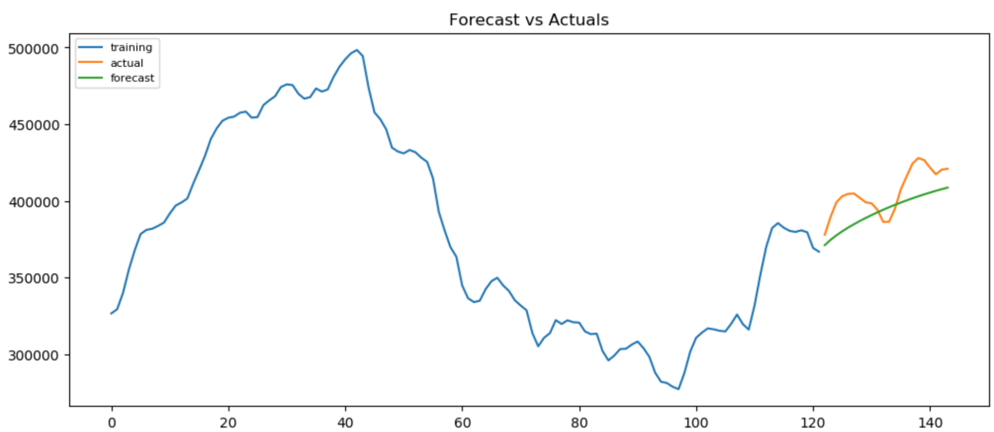


   The RMSE for TES with additive trend and seasonality is quite low when we do not damp the trend: 2479.56

3. SARIMAX
   SARIMAX produces a much better result than SARIMA as expected, but it didn't beat TES with or without damping. The RMSE for SARIMAX is 56,044.06

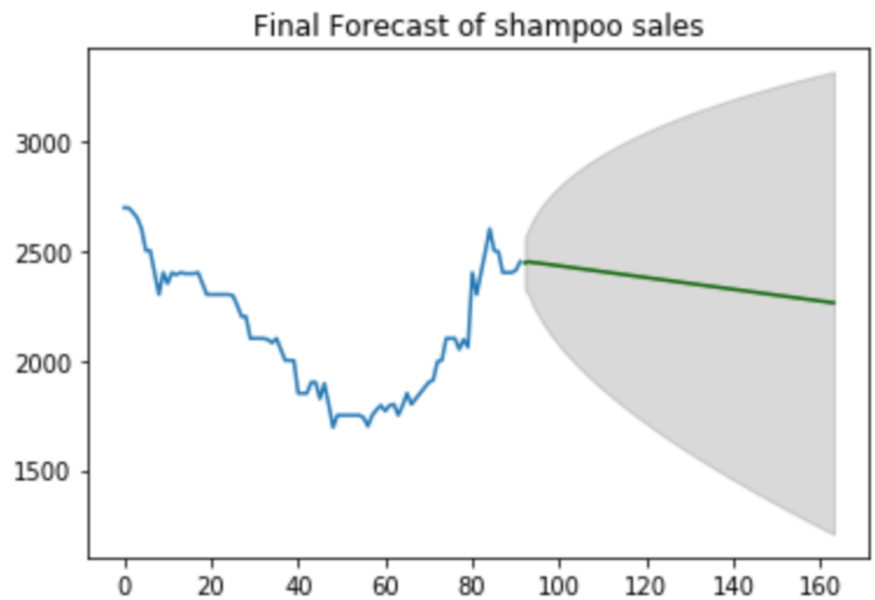


4. VAR
   Since VAR makes use of past features, we decided to impute the missing values in the median rental price.
   a. Imputation with the median

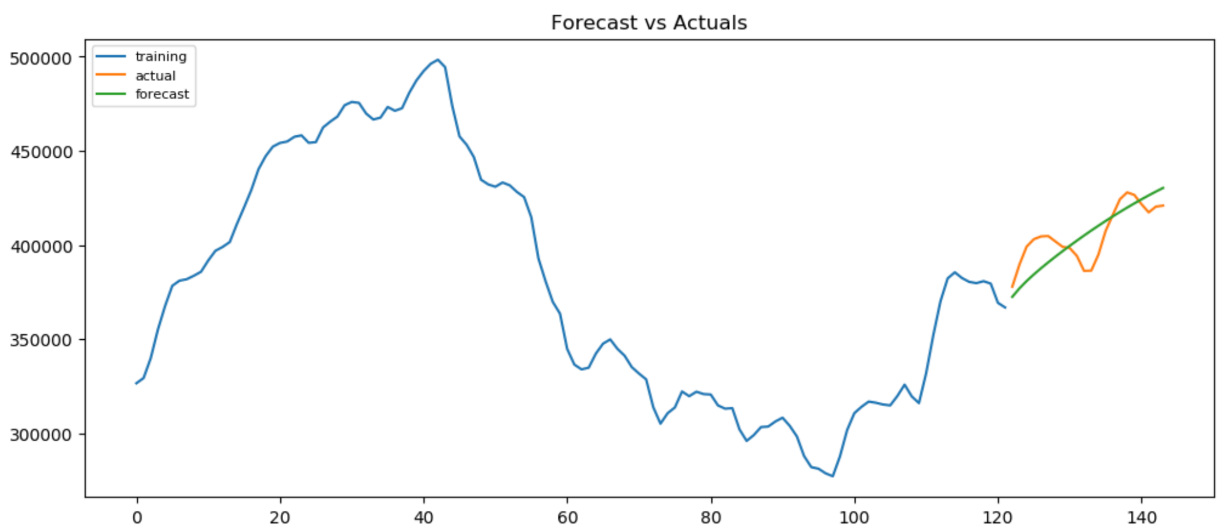


The RMSE is 15,571.23, which is much better than SARIMAX

   b. Imputation with ARIMA
      Since the median rental price has its own trend and possible seasonality, we tried to reverse it in time and use ARIMA to predict the beginning missing value. Here is what it looks like (after inverse).



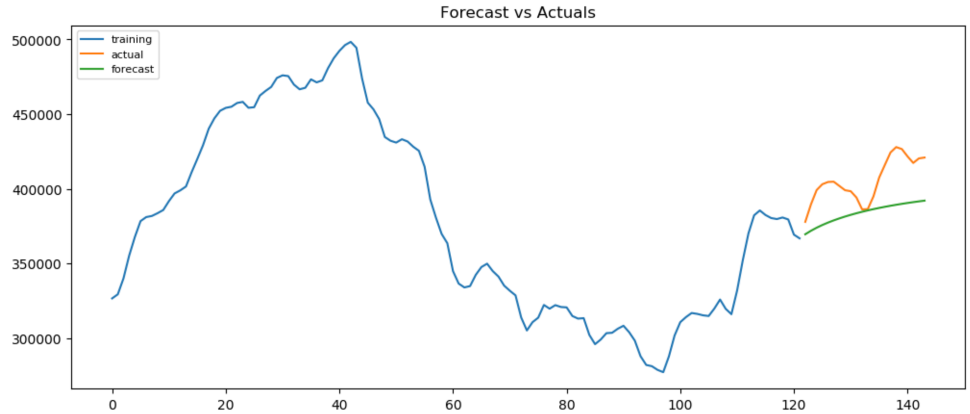
After imputation, we predicted the validation time range with VAR again:



The RMSE is 11,538.46, much better than imputation with the median value.

c. Imputation with VAR
   We also tried VAR to impute the missing rental price, however it did not perform

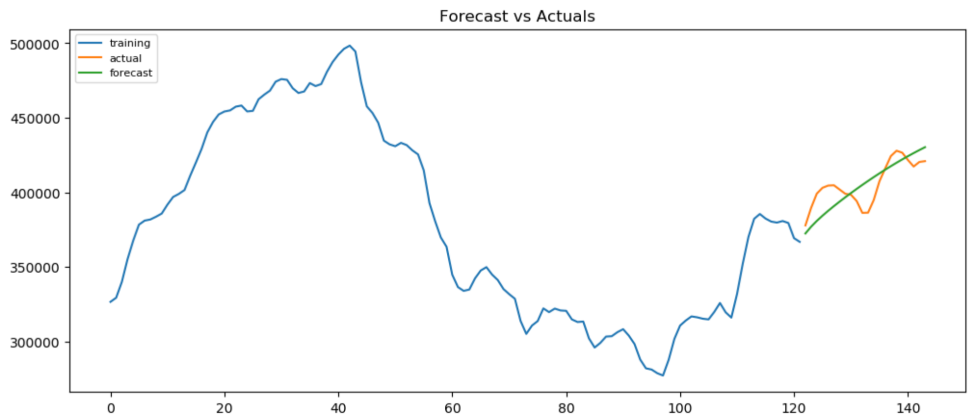
as well as the other models.



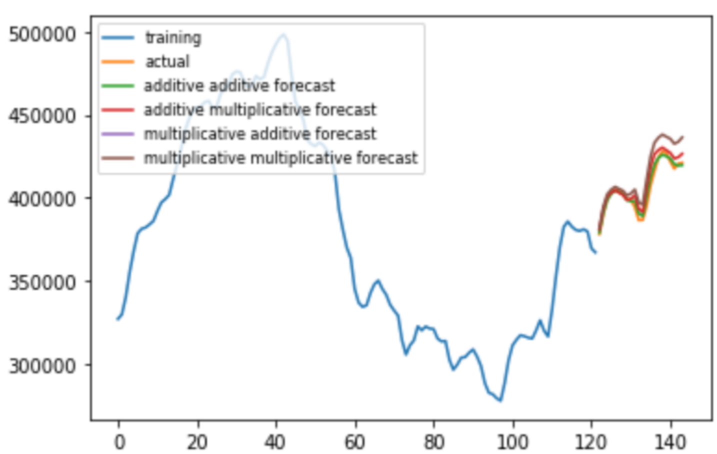
The RMSE for it is 24,313.45

3. Founding

1. VAR seemed did not capture the seasonality very well



2. TES predicts very well in the short-term



4. Forecasting

We decided to combine the prediction of TES (non-damp) and VAR with imputation from ARIMA because VAR doesn't capture the seasonality very well and TES might overestimate the trend. We trained on all data available and averaged the predictions of these two methods.