UNIVERSITY OF SAN FRANCISCO

CHANGE THE WORLD FROM HERE

MSDS 604

Time Series

Final Project

Report

Group Members:

- David Lyu
- Surbhi Prasad
- Vishwas Prabhu

UNIVERSITY OF SAN FRANCISCO

CHANGE THE WORLD FROM HERE

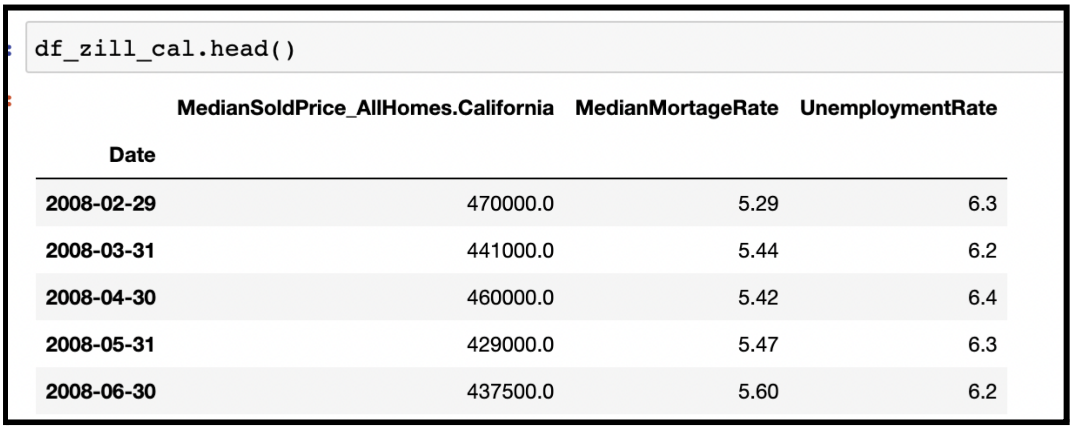
Table of Contents

Introduction

The zillow data for monthly median house price is available from 2008 to 2016 of which 2016 is kept as hold out test data to report performance of the final model. Other variables available are mortgage rate and unemployment rate which will be explored if they impact the house price.
Our aim is to predict the 2016 (test) sold price with the best selected model from candidate models.

Basic EDA

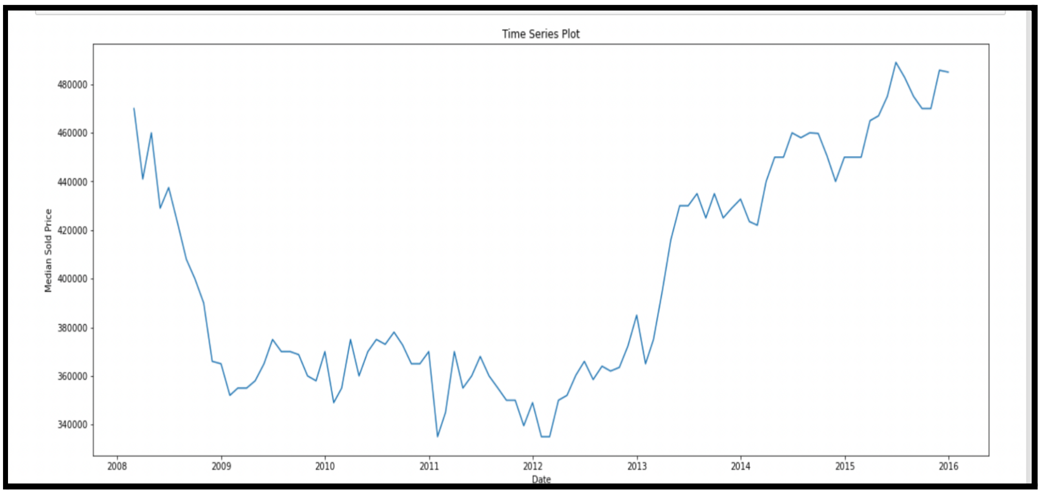
1. Basic checks- This data has no nulls or outliers as such and is directly used for further analysis.Data is monthly with 95 rows for 2008 to 2015.



```
df_zill_cal.head()
```

	MedianSoldPrice_AllHomes.California	MedianMortageRate	UnemploymentRate
Date			
2008-02-29	470000.0	5.29	6.3
2008-03-31	441000.0	5.44	6.2
2008-04-30	460000.0	5.42	6.4
2008-05-31	429000.0	5.47	6.3
2008-06-30	437500.0	5.60	6.2

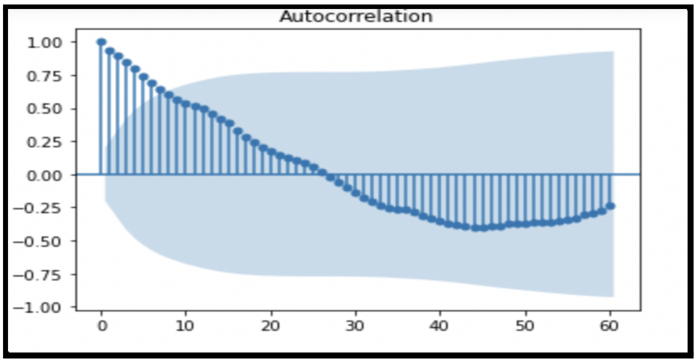
2. Time Series plot of Median Sold House price for train data

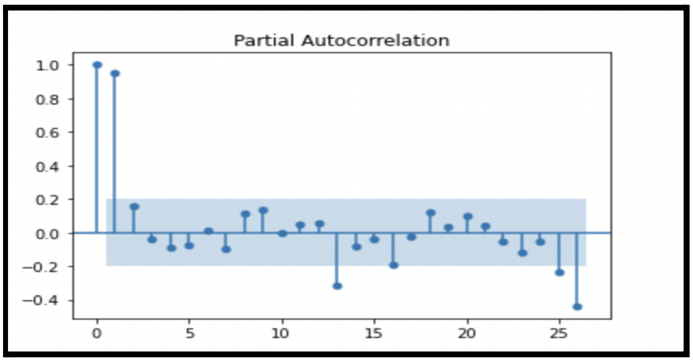


Looking at the Time series, the series doesn't look stationary, there is a slight increase in prices , so there is a base trend . We can't comment on seasonality by looking at plots though.Next step is to substantiate the same with an ADF test for stationarity.

Time Series- Additive or Multiplicative

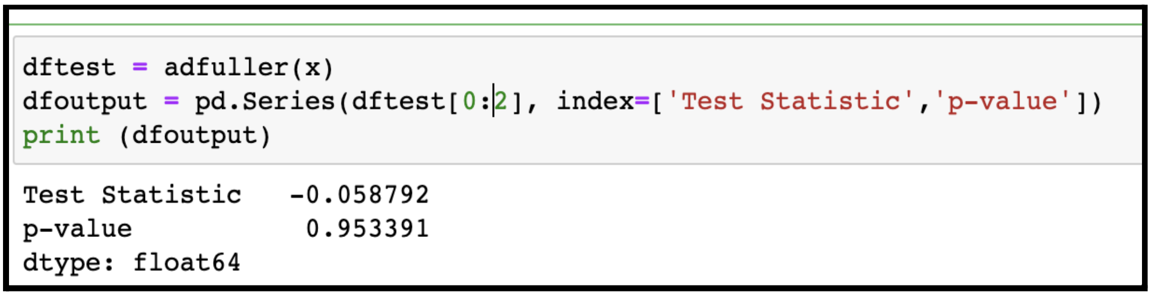
As we can see from the time series plot, there is no definite seasonal increase in amplitude of variation in time series with time. Hence, we will be considering time series to be additive. This can also be seen from below seasonal decomposition chart(next page) in which residuals have become stationary.





Test Stationarity and Detect Trends and Seasonality

3. ADF test for Stationarity of SoldPrice

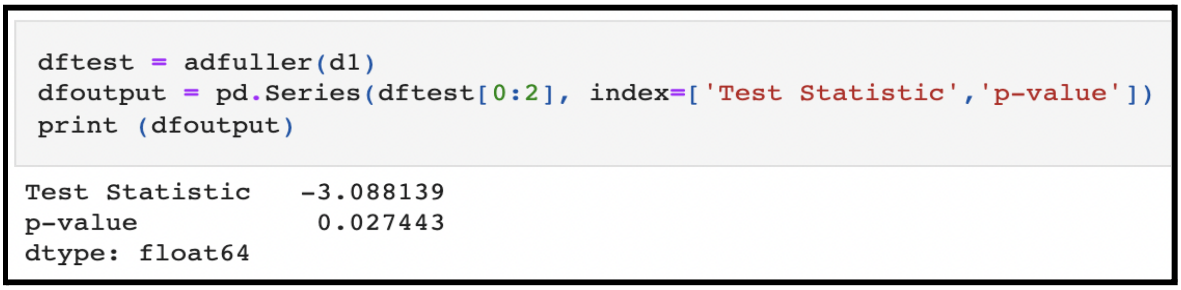


```
dftest = adfuller(x)
dfoutput = pd.Series(dftest[0:2], index=['Test Statistic','p-value'])
print (dfoutput)
```

```
Test Statistic   -0.058792
p-value           0.953391
dtype: float64
```

4.

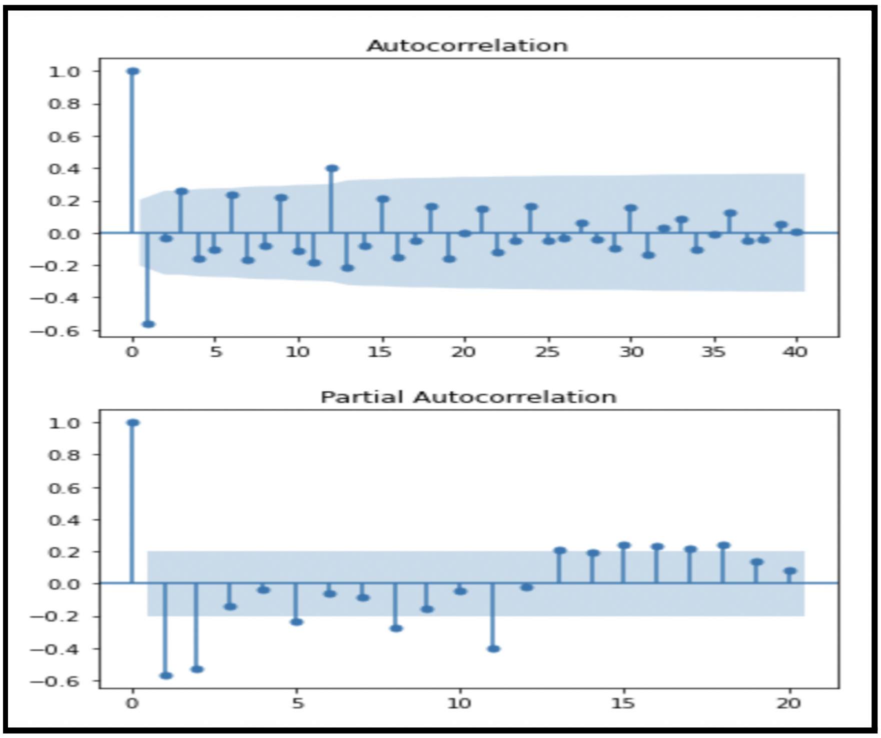
From the ADF test, as p-value >0.05 we can’t reject null hypothesis . Hence, we can't say the series is stationary. Hence the difference will be the same. So, Time series is differenced once to find P and MA from ACF and PACF plots if possible and recheck for stationarity from ADF which gives p-value less than 0.05 .

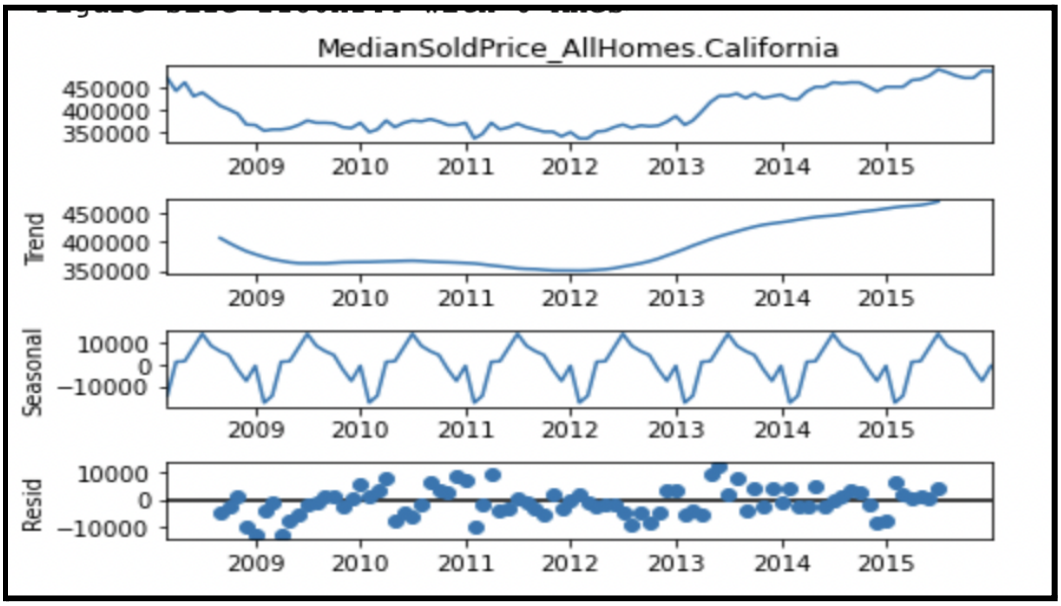


```
dftest = adfuller(d1)
dfoutput = pd.Series(dftest[0:2], index=['Test Statistic','p-value'])
print (dfoutput)
```

```
Test Statistic    -3.088139
p-value            0.027443
dtype: float64
```

Plots after difference, d=1 shows both plots shutting off which might indicate low orders of p/q=(0,1,2) . We can also see slight peaks in correlation after every 12 lags.
But can't say with confidence, So we tried the seasonal_decompose function to decide for itself in which we can see in the ACF plot also that there is seasonality(correlation increases every 12 months), it repeats every year=12 months. So, our few candidate models can have m=12 seasonality. This is somewhat reasonable since the house prices may vary in a seasonal pattern every year or every 12 month.

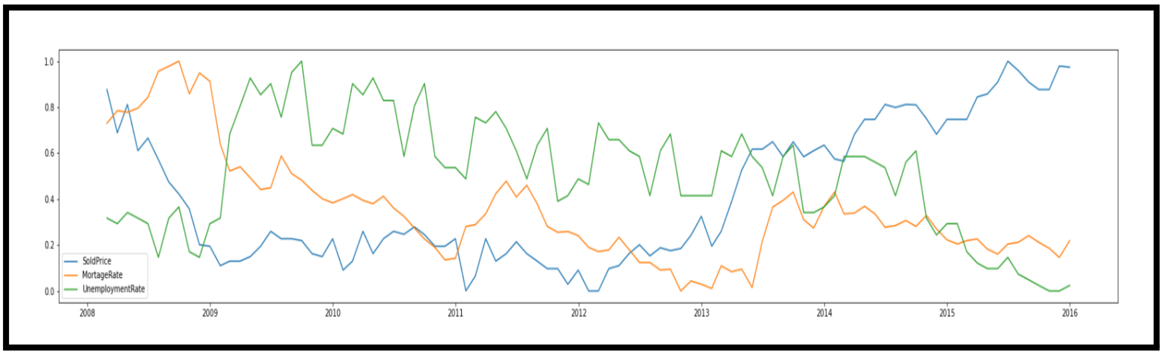


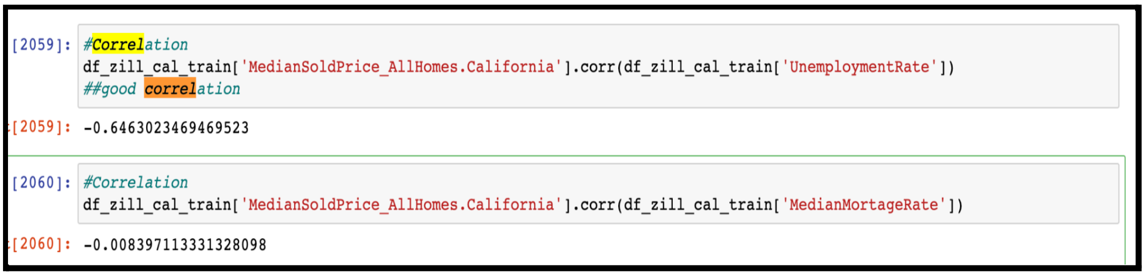


Since, we can't see define p also from the autocorrelation plot , we can try ETS too. We will skip the Prophet model as it won't work in this type of data where there are not strong seasonality or patterns.
Hence, we can select **(S)ARIMA and ETS** univariate methods.

Relation with Other Variables

We can check if all three variables impact each other as well as if Unemployment and Mortgage Rate impacts MedianSold Price scaled plots and correlation values. We can say it should be a one way causal relationship as Unemployment rate and Mortgage Rate can impact Selling Price of house while change in house price may not cause impact on employment /mortgage rates.Still can try VAR. We can see from correlation Unemployment seems to be correlated with prices.



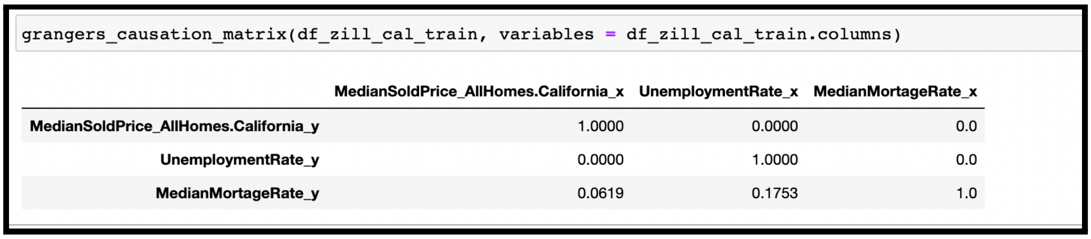


```
[2059]: #Correlation
        df_zill_cal_train['MedianSoldPrice_AllHomes.California'].corr(df_zill_cal_train['UnemploymentRate'])
        ##good correlation
[2059]: -0.6463023469469523

[2060]: #Correlation
        df_zill_cal_train['MedianSoldPrice_AllHomes.California'].corr(df_zill_cal_train['MedianMortageRate'])
[2060]: -0.008397113331328098
```

From plot as well as casualty test(p-values <0.05 for most variable combinations), we can select multivariate models also for prediction, **SARIMAX and VAR.**
For SARIMAX we will be choosing 3 candidate models based on both Unemployment rate and median mortgage rate, standalone unemployment rate and finally standalone mortgage rate as exogenous variables. The best model for each of these candidates will be selected based on the auto arima function which takes in scaled values of both the regressor and the exogenous variable.

We checked variables causation to each other for VAR using Grangers causation p-values matrix. As it's visible median sold price is impacted by unemployment but not mortgage(>0.05 p-value)



```
grangers_causation_matrix(df_zill_cal_train, variables = df_zill_cal_train.columns)
```

	MedianSoldPrice_AllHomes.California_x	UnemploymentRate_x	MedianMortageRate_x
MedianSoldPrice_AllHomes.California_y	1.0000	0.0000	0.0
UnemploymentRate_y	0.0000	1.0000	0.0
MedianMortageRate_y	0.0619	0.1753	1.0

Methods Descriptions

Univariate Models

A time series that consists of single (scalar) observations recorded sequentially over equal time increments.

MultiVariate Models

A Multivariate time series has more than one time-dependent variable. Each variable depends not only on its past values but also has some dependency on other variables.

Models Chosen

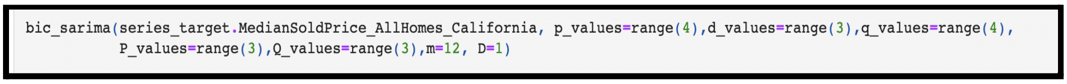
Based on above findings, we propose selecting best model from these methods:

1. Univariate - ARIMA(without seasonality), SARIMA
2. Multivariate models -SARIMAX (exogeneous impact) and VAR(endogenous vars)

We will select candidate models and best models from these methods based on lowest RMSE from one step cross validation function.

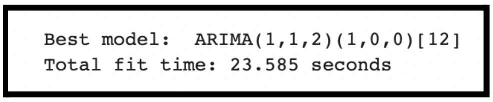
Models Selection Criteria and Methods

1. For ARIMA models, we tried a range of p,d,q and P,Q with D=1 and picked candidate models with **lowest bic scores.**



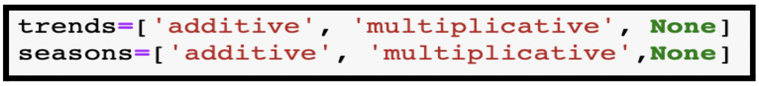
```
bic_sarima(series_target.MedianSoldPrice_AllHomes_California, p_values=range(4),d_values=range(3),q_values=range(4),
          P_values=range(3),Q_values=range(3),m=12, D=1)
```

2. SARIMAX- three types of models were tried. Model with Mortgage rate only, unemployment only and model with both. The best model was chosen based on auto.arima **(Auto ARIMA takes into account the AIC and BIC values generated (as you can see in the code) to determine the best combination of parameters.)** keeping seasonality as 12.



```
Best model:  ARIMA(1,1,2)(1,0,0)[12]
Total fit time: 23.585 seconds
```

3. For ETS model: we tried all combinations for trends and seasonality(12)- and selected based on **lowest RMSE**.



```
trends=['additive', 'multiplicative', None]
seasons=['additive', 'multiplicative',None]
```

4. Prophet Model- We didn't choose as the data was not suitable for such a model with not strong seasonality(weak) patterns.
5. VAR models- lag =11, 12 (based on **lowest AIC**) were tried and the lowest RMSE model was selected.

Validation Method

We have used one step forward validation as it is a time series model. We have divided our train data (upto Dec'15) in 80:20 ratio and utilising one step forward forecast method for validation to select the best model.

Candidate Models from each method

Four types of Model:

We will be choosing models based on the following criteria:

1. ARIMA
   a. (0, 1, 0), (0, 0, 0, 0)
   b. (0, 2, 1), (0, 0, 0, 0) - selected based on bic
2. SARIMA- (0, 2, 1), (0, 1, 2, 12) - selected based on bic
3. SARIMAX-
   a. (1, 1, 2), (1, 0, 0, 12)
   b. (0, 1, 0), (0, 0, 0, 0) - based on auto arima.
4. VAR lags=12 & 11 -selected based on aic/bic

Finally, these models are compared based on lowest RMSE as a metric on validation set to select one best final model.

Comparison of Models RMSE

Candidate	Type	Variables	Exo/Endo Vars	Model HyperParameters	Validation RMSE	NRMSE*
1	ARMA	Univariate	None	(0, 1, 0), (0, 0, 0, 0)	8,162	1.75%
2	ARMA	Univariate	None	(0, 2, 1), (0, 0, 0, 0)	10,324	2.22%
3	SARIMA	Univariate	None	(0, 2, 1), (0, 1, 2, 12)	12,599	2.71%
4	SARIMAX	Multivariate	Unemployment Rate	D=0 (1, 1, 2), (1, 0, 0, 12)	9,086	1.96%
5	SARIMAX	Multivariate	Unemployment Rate	D=1 (1, 1, 4), (0,1,2, 12)	10,228	2.20%
6	SARIMAX	Multivariate	Mortgage Rate	(1, 1, 2), (1, 0, 0, 12)	9,879	2.12%
7	SARIMAX	Multivariate	Both	(1, 1, 2), (1, 0, 0, 12)	9,220	1.98%
8	ETS	Univariate	None	Multiplicative Trend , No season	9,066	1.95%
9	Prophet	Univariate	None	Seasonality=12 added	16,936	3.64%
10	VAR	Multi Variate	Unemployment	Lag=11	27,268	5.86%
11	VAR	Multi Variate	unemployment	Lag=12	26,928	5.79%

NRMSE is normalised RMSE with mean of validation data sold price.

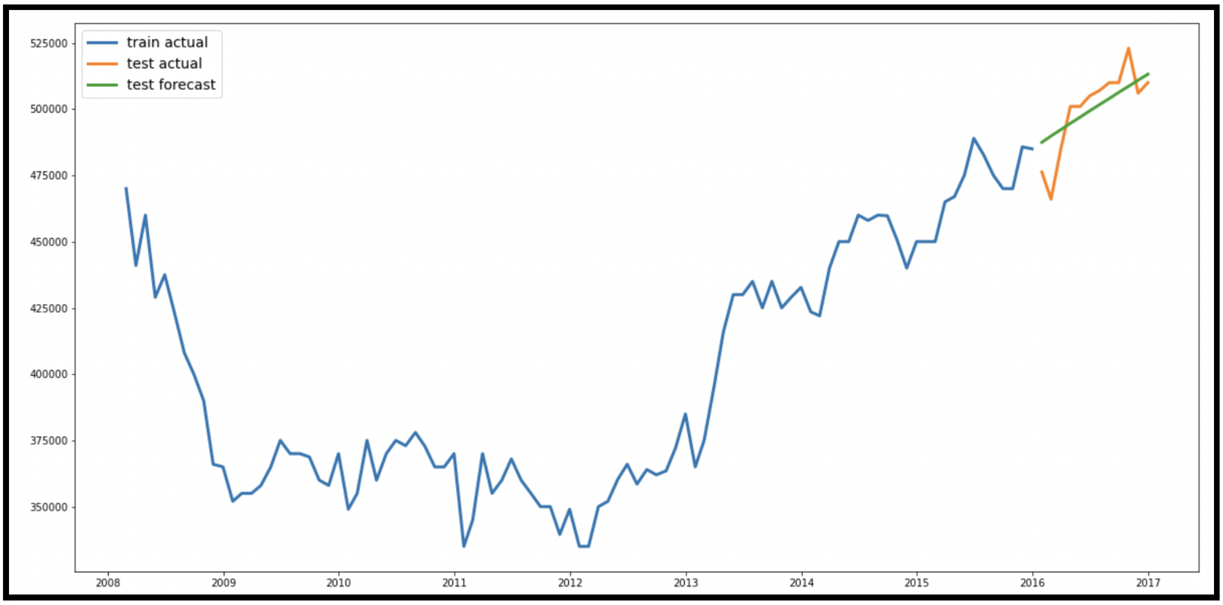
Decision between ARMA and ETS

Though ARIMA is giving lowest RMSE based on one step forward forecast validation method, we know that ARIMA(0,1,0) model will give a constant line which is good for short term but not good for long term forecasts. Hence, we have not considered the same as one of our candidates for final comparison.

Hence, our lowest RMSE model i.e. best model is the ETS (multiplicative trend, no seasonality) model.

Final Selected Model

Model: ETS (multiplicative trend, no seasonality).
Test dataset RMSE: 9,813.03 dollars
Test NRMSE: 2.11% (RMSE normalised by test mean sold price)


train actual
test actual
test forecast