Time Series Analysis with Zillow Dataset

Group 21: Hashneet Kaur, Huidong Xu, Shuyan Li

1. Introduction

In this project we have analyzed the California Zillow housing dataset to do the time series analysis of the median sale prices, i.e to see is there is a trend/seasonality in the housing sale prices. We have also found the best time series model to predict the housing sale price at a given point of time.
We have split the data into a training set and a validation set for finding the best model that can predict the median selling prices for all houses in California at a particular time. We have trained our models on the training data using cross validation and found the best model based on it's **RMSE - Root mean square error** for the predictions it does on the validation set.

For our analysis we have tried different models:

1. Seasonal Autoregressive Integrated Moving Average(SARIMA)
2. Exponential Smoothing
3. Univariate Prophet
4. Multivariate Prophet
5. Seasonal Autoregressive Integrated Moving Average with external variables(SARIMAX)
6. Long short-term memory(LSTM)

To find the best model from these, we used cross validation with split size 0.67.
We used **RMSE - Root mean square error** as the metrics to compare different models and find the best one.

Approach Followed:

For each category of model we tried 2,3 different parameters with a test train and validation set which gave us 2-3 different models for each algorithm. From these we selected the ones with lowest RMSE. This gave us different candidate models.

So we had 6 different candidate models, one from each algorithm. From these we selected the one with the lowest **RMSE as our final model.** We then used this final model to check for the RMSE on the test set.

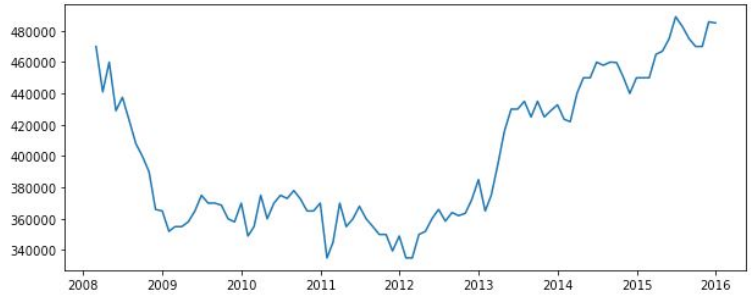
2. Preprocessing

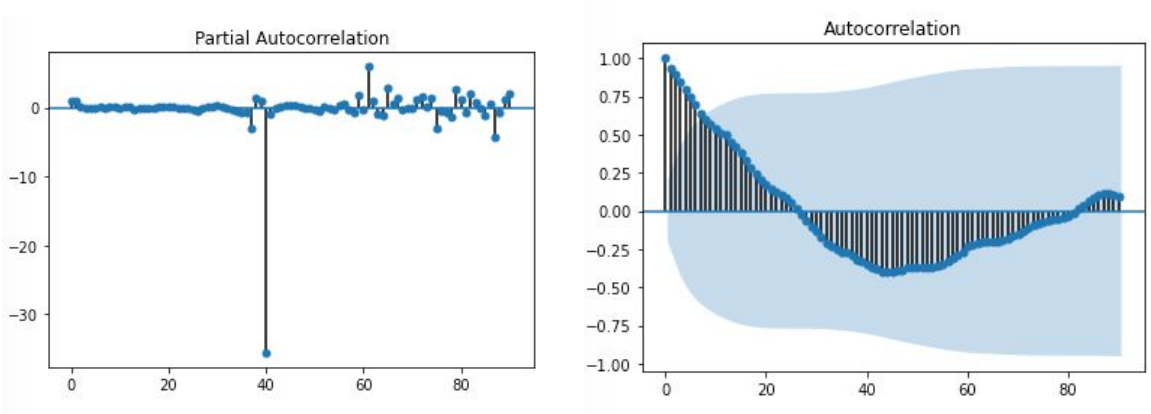
For the preprocessing part, we split the dataset into train and validation set by the date '2014-12-31'. The training dataset contains observations before 2014-12-31(included) and the

validation dataset includes the data after 2014-12-31(not included). For cross validation we used the split size as 0.67.

The test set has the data after ‘2016-01-31’ to ‘2016-12-31’.

Analysis of the dataset with plots - data trend, ACF and PACF plots.





From these plots we analyzed that the data has a trend and seems to have a seasonality of 12 months. The data is also not stationary. To confirm our analysis about the data being stationary we also performed the ADF test. The ADF test p-value is 0.95 which is larger than 0.05, so the data is not stationary.

3. Model Fitting

In this part, we fit six kinds of models in total, including SARIMA, ETS, Univariate Prophet, Multivariate Prophet, SARIMAX, and LSTM. Then we select the best model by evaluating each model’s RMSE of actual values and the model prediction.

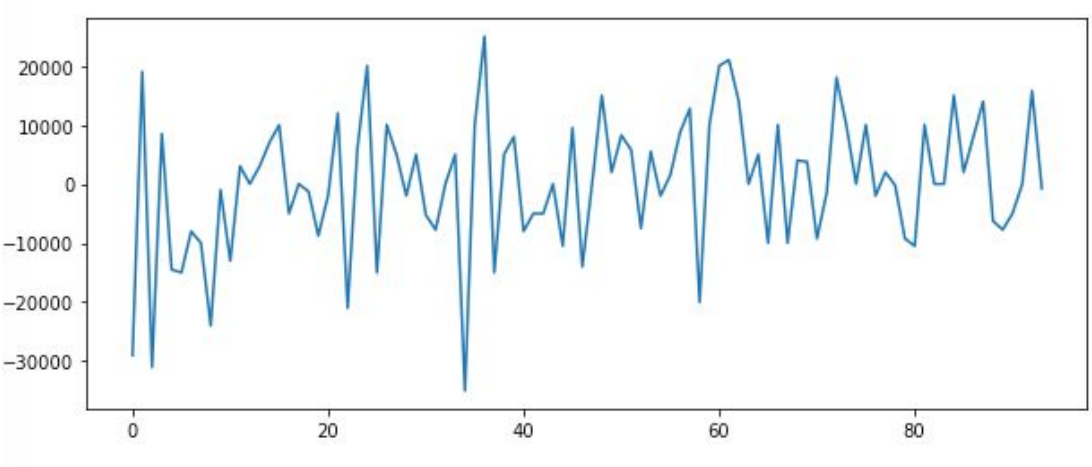
1. SARIMA Model

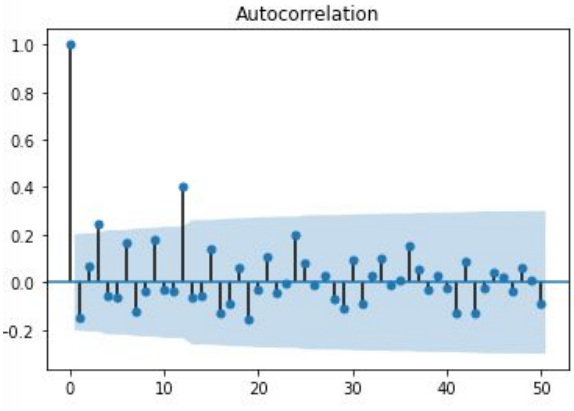
1.1. Hyperparameter Tuning

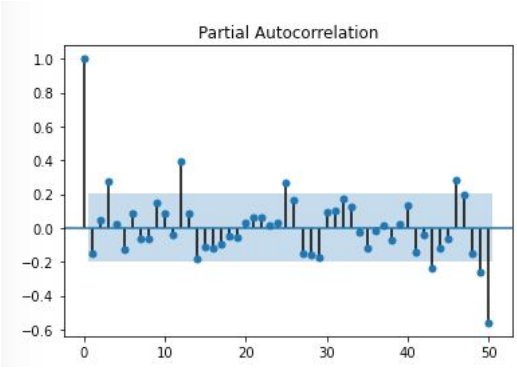
SARIMA has 6 components - trend components of autoregression order(p), differencing order(d), and moving average(q) order as well as that of seasonal component(P, D, Q). We created a search space containing different values for these parameters and conducted cross-validation to each model.

To perform SARIMA, we difference the data by 1 for trend.

Plots after 1st differencing

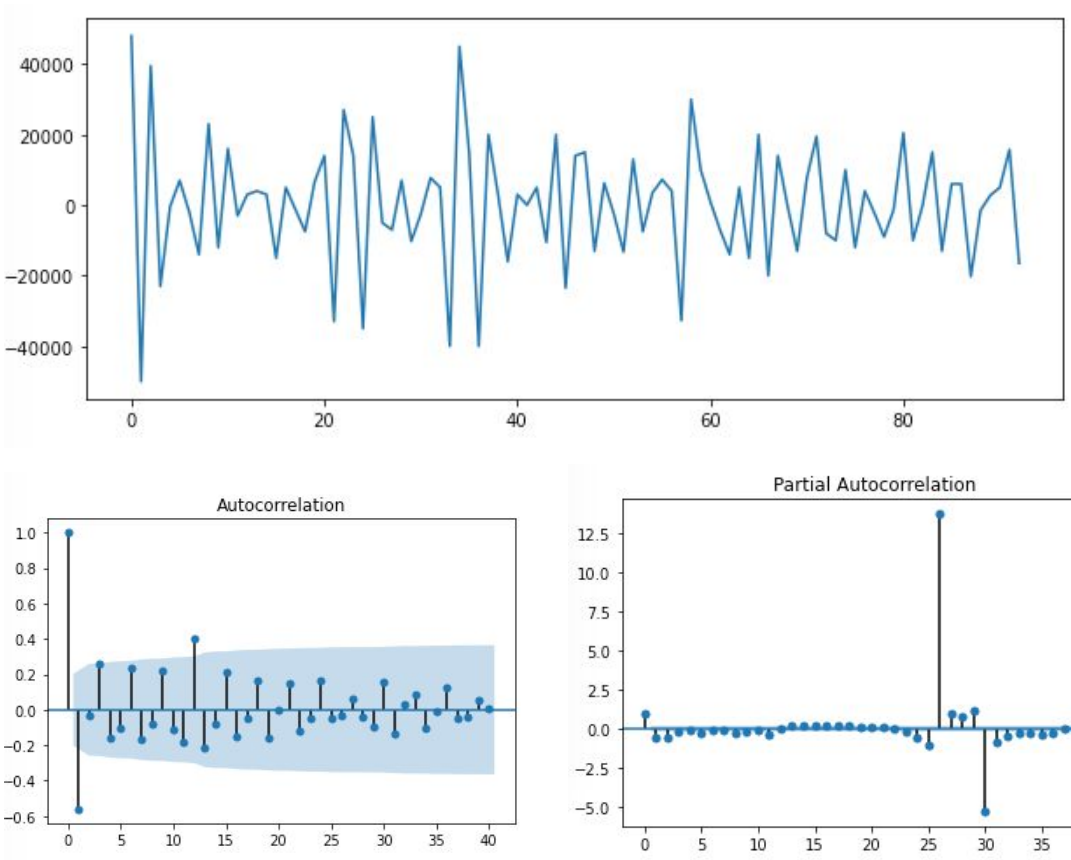






The plot looks more stationary but it may still be affected by the seasonal trend. The ADF test result is 0.027 which is ok but just in case, we decide to have data differencing one more time.

Plots after second differencing:



From the plot, we could see that d is 2 and data becomes stationary. The ADF test score is extremely small. As for the rest parameters, we set m as either 3 or 12 since there is possibly a quarterly trend or yearly trend. And for p, q, P, Q, we tune those parameters in the range of 0 to 6.

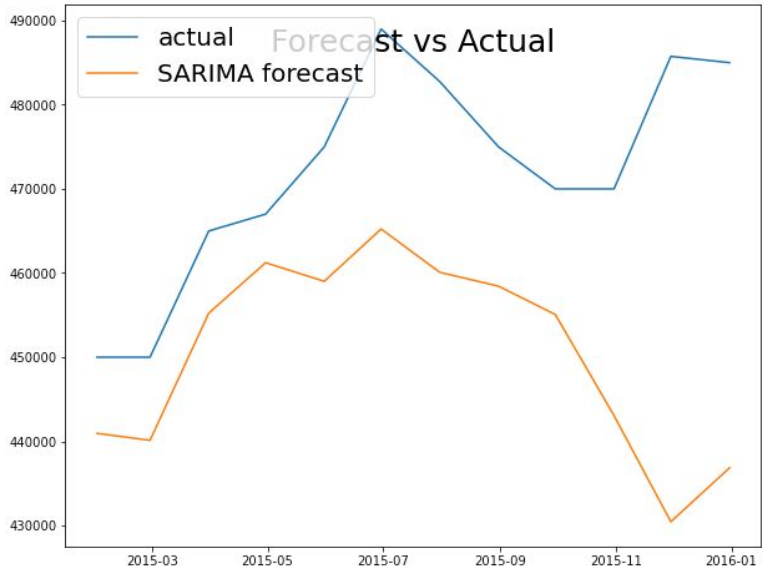
Cross-validation

As for time series cross-validation, it is different from the other kinds of the dataset. It will do a walking forward evaluation on the test dataset. We choose an initial training dataset and used it to fit a model then forecast one step as a prediction and keep adding it to the history data to predict the next step. We will select the lowest RMSE model as a candidate model of SARIMAX for final evaluation

The Best SARIMA Model:

Candidate 1: SARIMA(trend = (0,2,5), seasonal = (0,1,5,12)).

- The RMSE of cross-validation is 13976.35.
- Here is the plot of forecast and validation values(From 2015-01-31 to 2015-12-31)



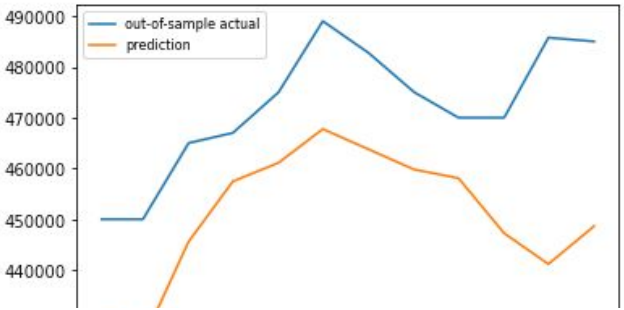
2. ETS Model

We choose to use ETS cross-validation with a split value of 0.67. We checked for different values for trend and seasonality - additive/multiplicative. The metric we are using is the RMSE. Also, we set the seasonal period as 12 to see the seasonal trend either quarterly or yearly.

Best ETS Model

Candidate 2: ETS(Trend="additive", Seasonal="additive", m=12, damped=True)

- RMSE of Cross-validation: 23576.18
- Prediction on the validation dataset.

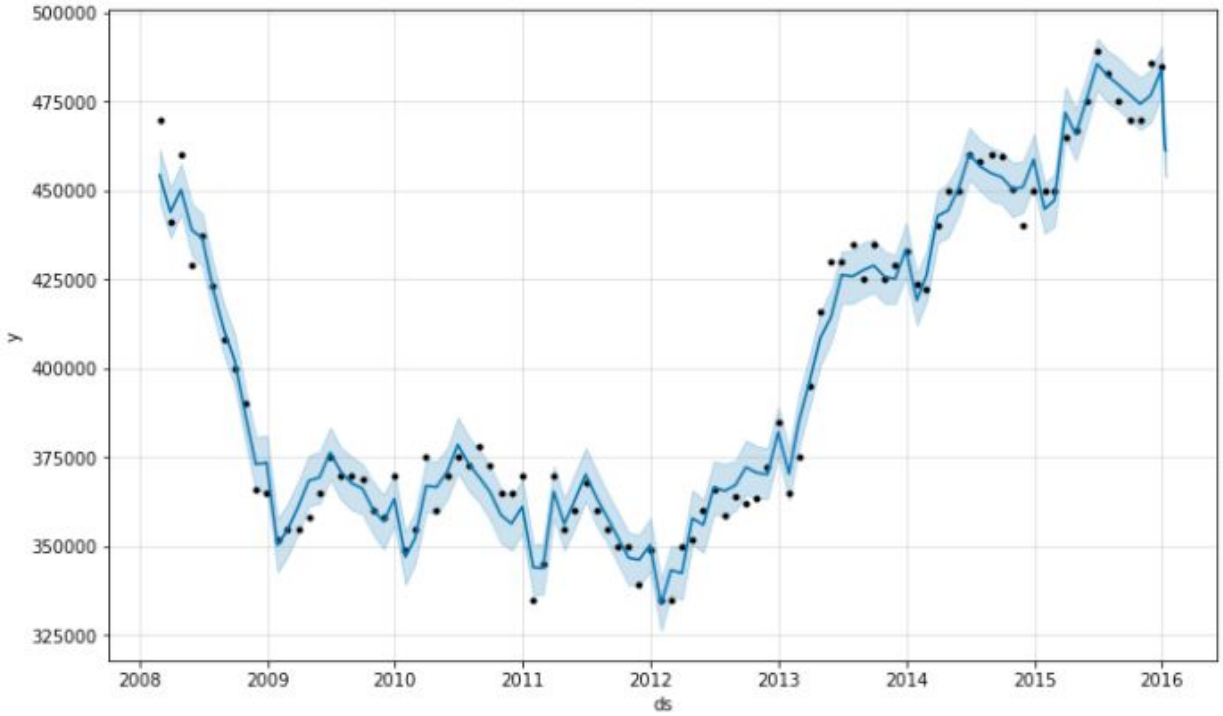


3. Univariate Prophet

The prophet model is fast and is robust to outliers, missing data, and dramatic changes in the time series.

We used a prophet model using the default parameters and fit the model with the training dataset. For the validation, we used the model to conduct the prediction and calculate the RMSE between the prediction and validation dataset.

Prophet plot for history data

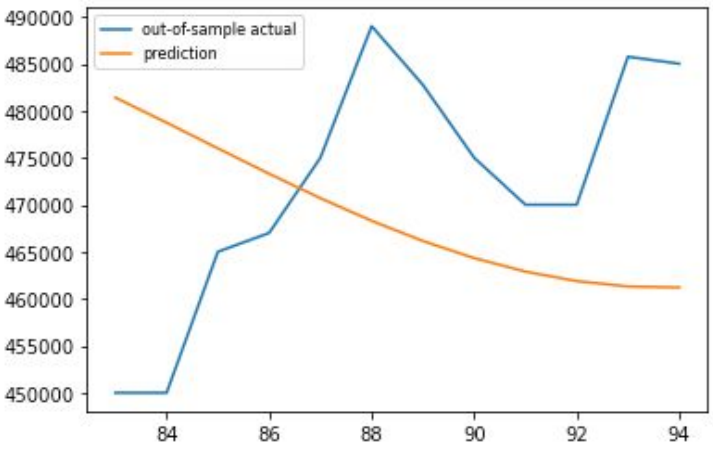


The Best Univariate Prophet Model

For the univariate prophet model, we only have one model and it is selected as the candidate model.

Candidate 3: Prophet()

- RMSE: 18460.03
- Prediction on the validation dataset.



Multivariate Models

For the next two models, we are choosing two multivariate time series models because, in the data frame, there are two other factors ‘Median Mortgage Rate’ and ‘Unemployment rate’. Since in the real world, the mortgage rate, as well as the unemployment rate, could have a negative effect on the affordability of a person or family to a house. So the price could be affected by these two factors. We hypothesized, in these two models, the mortgage rate and the unemployment rate will be exogenous variables to the model and used the Prophet and SARIMAX models.

4. Multivariate Prophet Model

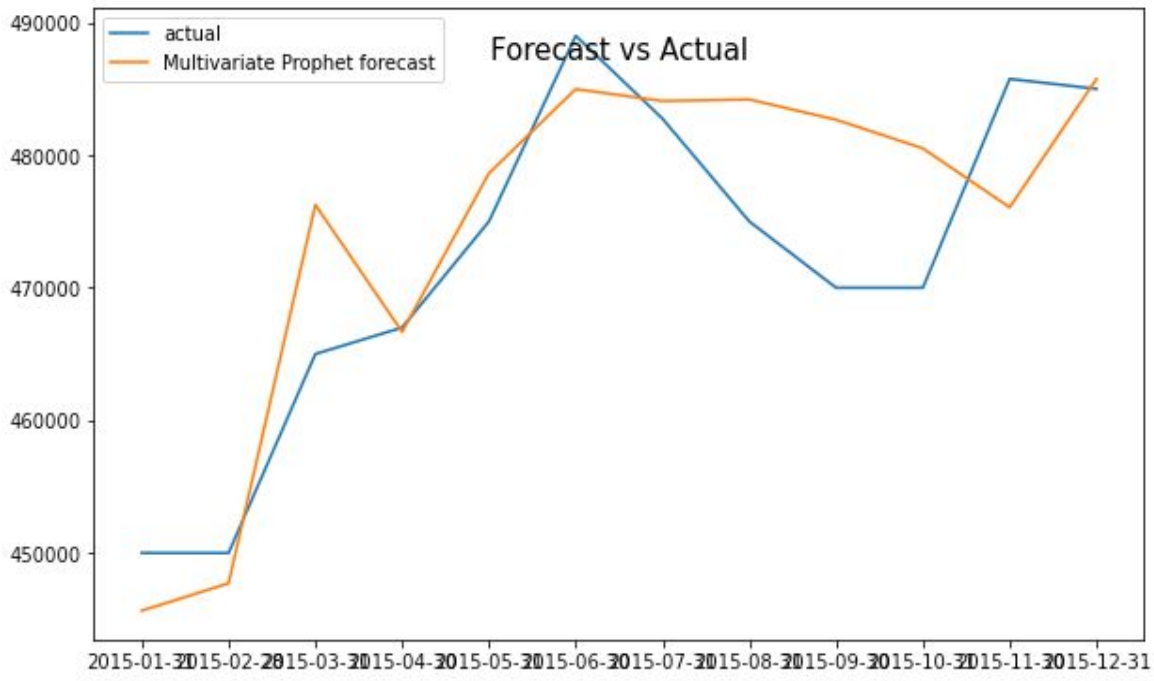
Along with the steps we followed in univariate prophet we will add a regressor to the prophet model. For this model, we add only the **‘median mortgage rate’** because we think this factor plays a more important role since it is more directly related to the house price.

The Best Multivariate Prophet Model

Candidate 4: prophet().add_regressor('mortage_rate')

- RMSE: 7257.67
- The plot on the prediction

Prediction plot for validation set



5. SARIMAX

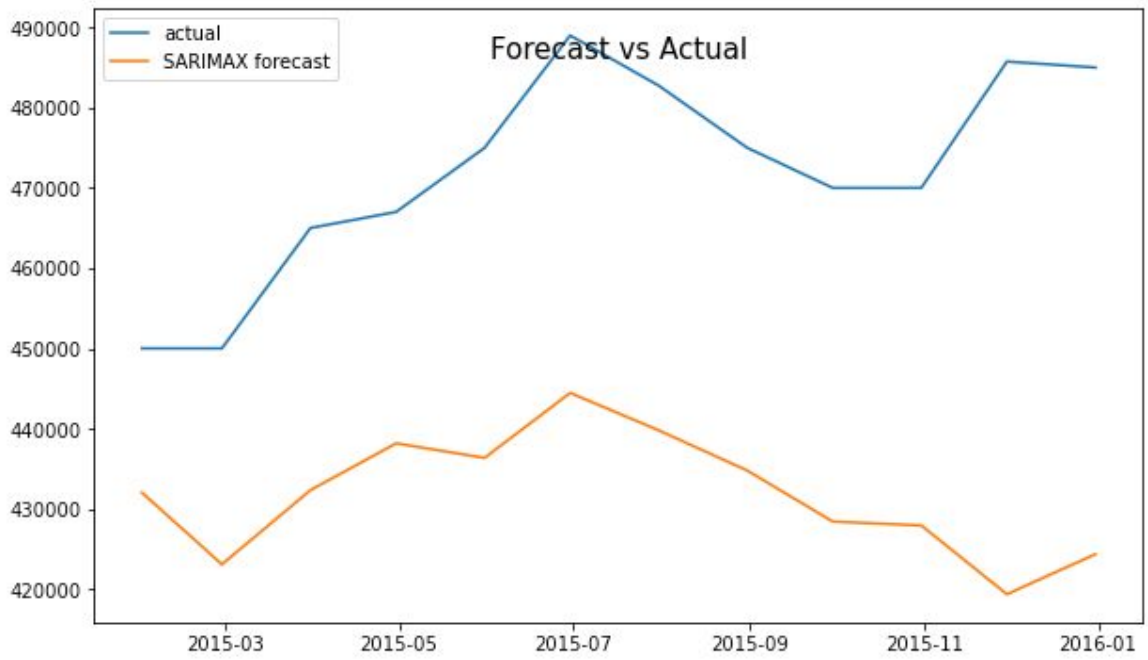
For the SARIMAX, similar to SARIMA, it considers the external variables and unlike the prophet, the SARIMAX contains more trend and seasonal components. Therefore, there is more space for us to find a better multivariate model with trend and seasonality components

We search for the best model using p in the range of 0-5, q in the range of 0-5, d in the range of 0-2. The seasonality components are D = 1, m = 12, P in the range of 0-3, Q in the range of 0-3. However, before the step of hyperparameter tuning, we normalize the data using Min Max Scalar.

The Best SARIMAX Model

Candidate 5: SARIMAX**(trend = (4,1,0), seasonal = (0,1,2) [12])**

- RMSE: 42303.21
- The plot on model validation



6. LSTM

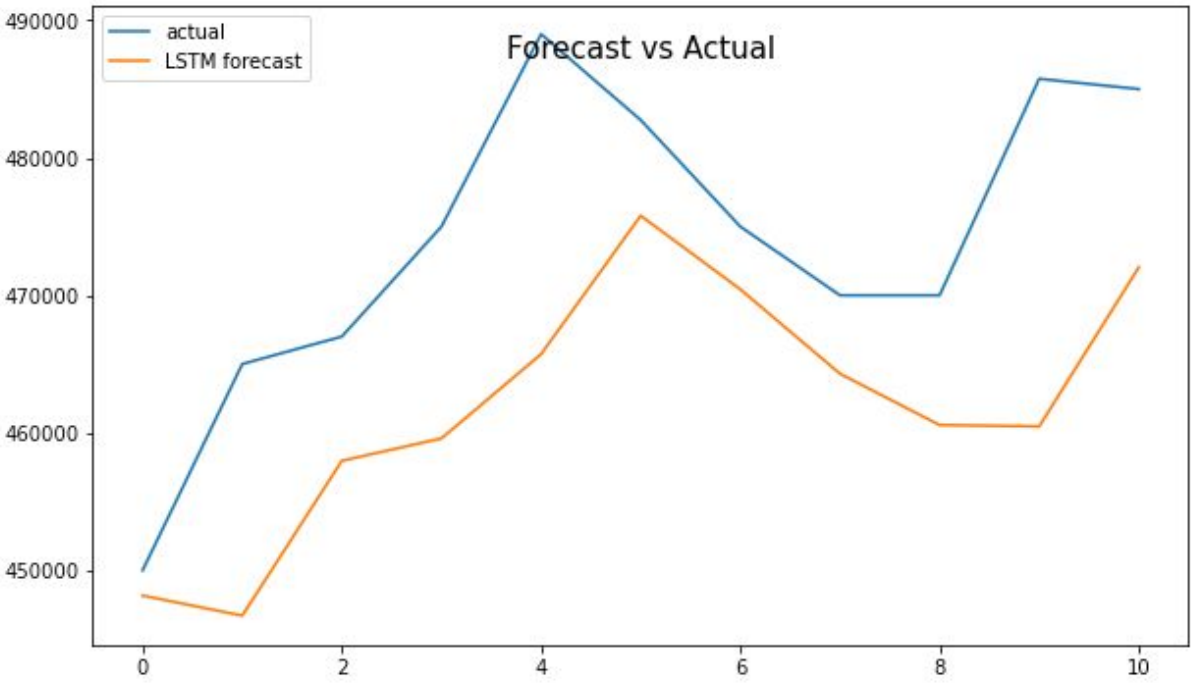
This is a model that the deep-learning LSTM applies to time series analysis. It will have the possibility with larger power in learning the context of the time series required for making predictions. Therefore, it is worthwhile to train this dataset there might be some hidden information that could be captured by the LSTM model.

For the preprocessing part, we will need to convert series to supervised learning required dataset format and also normalize them. Next, choose the predictors that we need to include. Then, we build the LSTM model, add a dense layer, and set the loss function as the mean absolute error.

Best LSTM Model

Candidate 6: LSTM()

- RMSE: 14133.32
- The plot on model validation



Candidate Model Summary:

Different candidate models selected and their RMSE on the validation dataset.

Models	**RMSE on validation set**
SARIMA(0, 2, 5), (0, 1, 5, 12)	26205.94
ETS(Trend="additive", Seasonal="additive", m=12, damped=True)	23576.18
Univariate Prophet()	18460.03
Multivariate Prophet	7257.67
SARIMAX(4,1,0)(0,1,2)[12]	42303.21
LSTM	14133.32

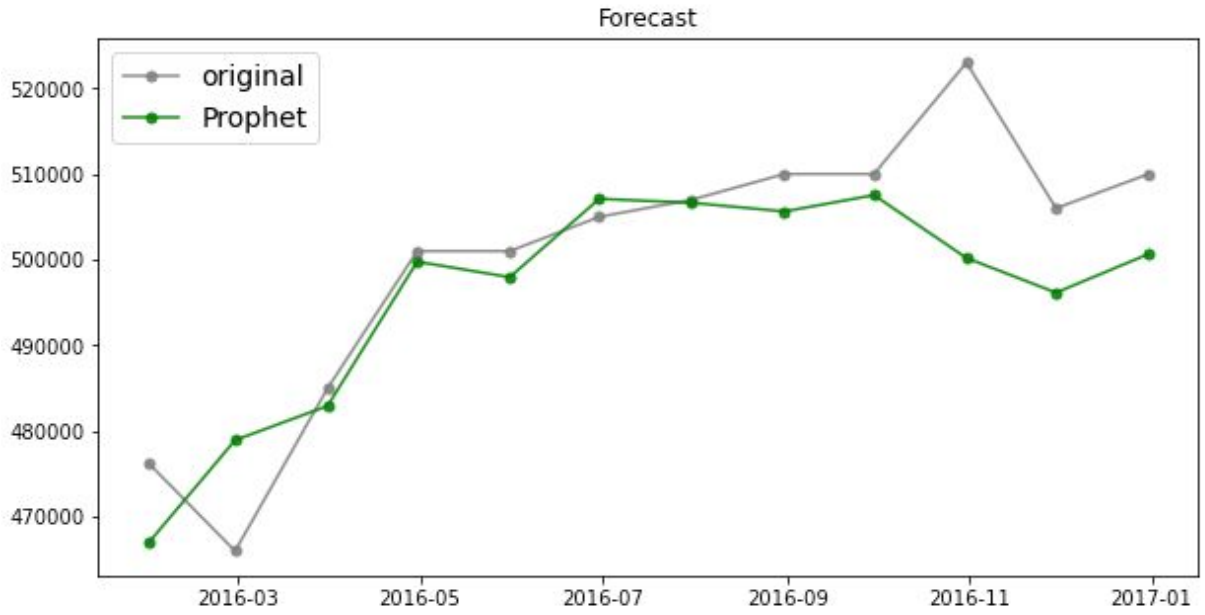
The Multivariate Prophet model with regressor Median Mortage Rate performs the best among all candidates. The RMSE score for this model on the validation set is 7257.67, so we pick this as our final model.

4. Final Model Evaluation

As our final step we will fit the final model we selected - Prophet with one regressor - to the test dataset.

The **RMSE** of our final model on the test dataset - **2016-01-31** to **2016-12-31**: **9140.50.**

Plot for test predictions using the final model - for data between 2016-01-31 to **2016-12-31**



From the above plot, we can see that our best model performs pretty well on predicting the median sold price for the test dataset, 2016-01-31 to 2016-12-31. The values predicted almost follow the trend of the actual values for the test dataset.

So as per our analysis using the California zillow dataset we can say that the median sale prices for the houses in california has a trend that follows over time and this trend in median sale prices is also strongly related with the variable median mortgage rate. So the best model to predict the median sale price at any particular time by using the **prophet model with median mortgage rate as the regressor.**