Predicting House Price of San Francisco

My Colab link

Overview

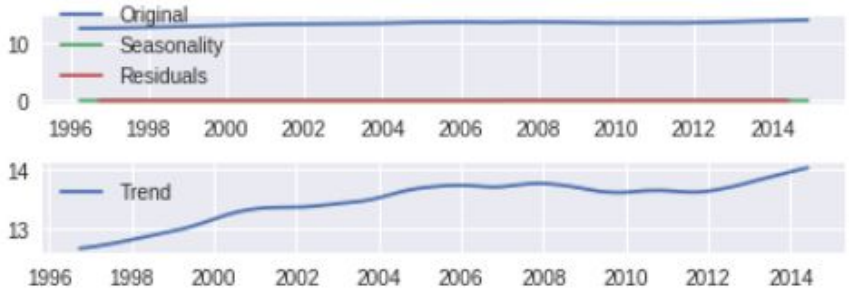
Prediction of house price is always a popular topic for the public. In this project, I utilize data from Github to predict the house price in San Francisco during 2015 and 2018 with two different models, time series model(ARIMA) and machine learning model (linear regression). By comparison of two models, we can have a more comprehensive understanding of house price in San Francisco.

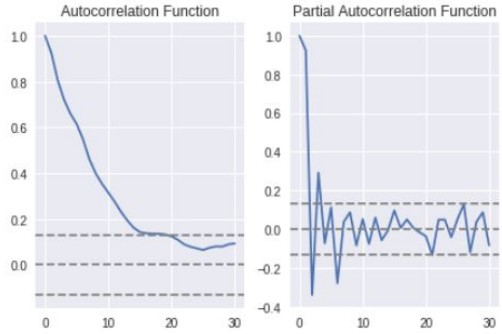
Technique

The original data has 23 years of American house price data. For both models, I select the last four years(48 months) of data as the testing dataset and the others as training dataset. In the time series model, I use ARIMA to do the predictions and in the machine learning model, I use linear regression. After comparing the predicted values with testing values, we can derive the accuracy of each model.

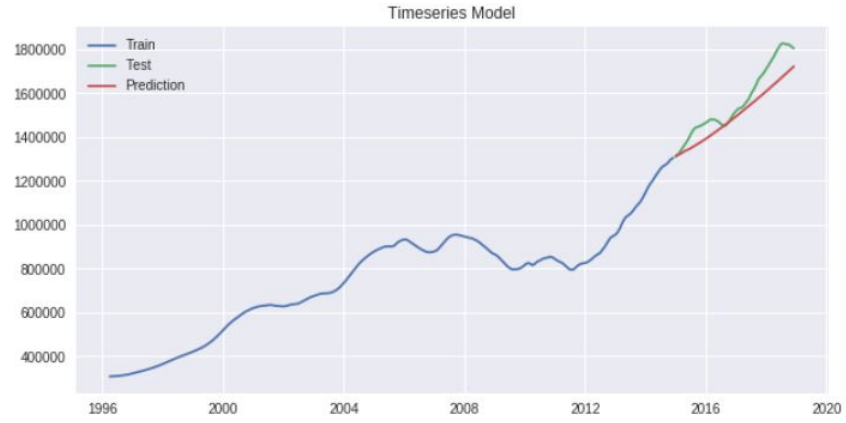
Model 1: Time series

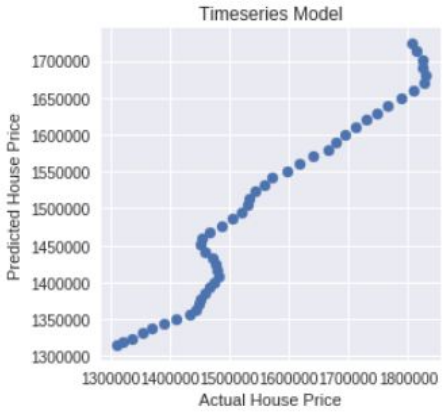
To establish an ARIMA model, firstly I need to test the stationarity of the data; secondly I do log transformation and time shift of data in order to remove the trend component and get stationary data; thirdly I lot ACDF graph to find the parameters *p* and *q* in ARIMA command.



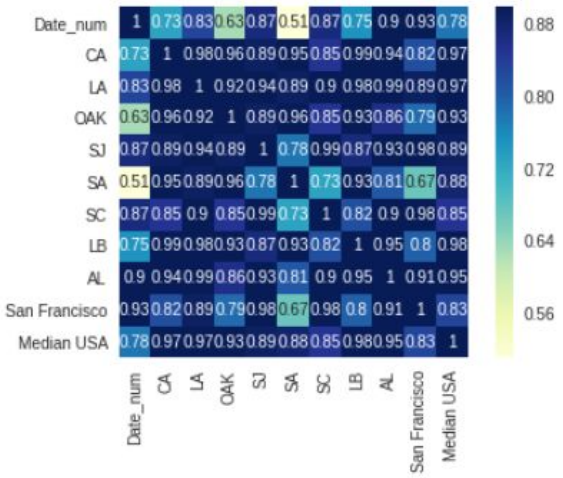


At last, I use training dataset to establish the ARIMA model and then use the model to predict the testing dataset. I plot the training dataset, testing dataset along with the predicted data in the left graph below and the comparison of testing values and predicted values on the right graph below.





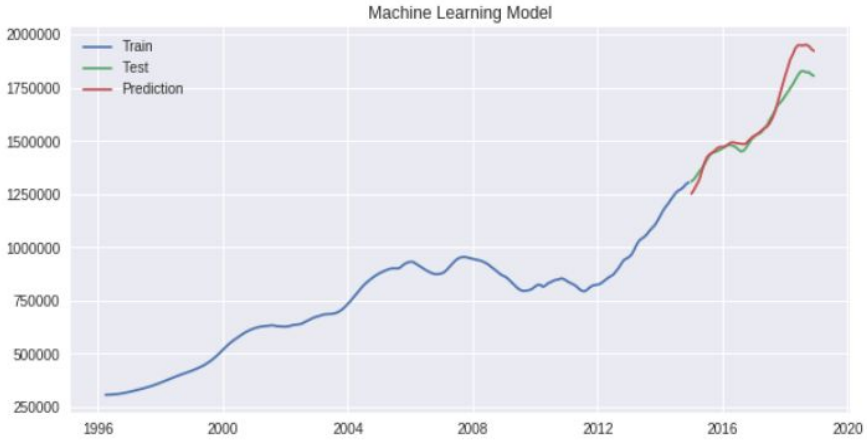
Model 2: Machine Learning

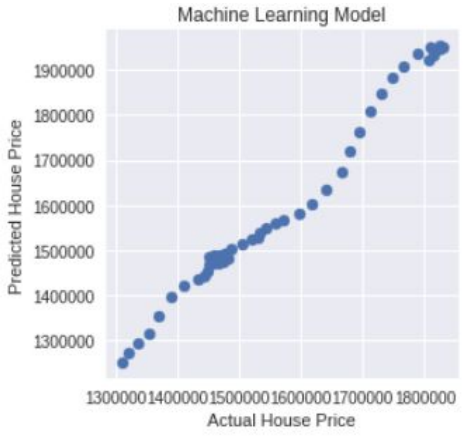


To establish a linear regression model, firstly I select the possible features I think will influence house price in San Francisco. Possible features include house price of USA, house price of California, house price of seven biggest cities in CA such as Los Angeles, Oakland and Sacramento, and also the numerical form of date named 'Date_num'.

Secondly, I plot a heatmap of features and target in the right graph and select all features that have a high correlation with the target as the independent variables.

At last, I use training dataset to establish the linear regression model and then use the model to predict the testing dataset. I plot the training dataset, testing dataset along with the predicted data in the left graph below and the comparison of testing values and predicted values on the right graph below.





Conclusion

From the graphs of two predicted results above, we can intuitively see the ARIMA model better fits the original data in the last two years(2017-2018) while the machine learning model better fits the original data in the first three years(2014-2016). To draw a more precise conclusion on which model is better, I calculated the Accuracy for both and the corresponding values of the ARIMA model and the linear regression model are 0.75 and 0.77 respectively. In this case, we can say that linear regression outperforms linear regression model slightly.