# Forecasting U.S. Energy Review

**Introduction:**

This timeseries dataset regarding the retail motor gasoline and on-highway Diesel fuel prices had some noticeable issues in terms of supplying an accurate forecast model for this assignment. With a range of 26 years, there were two separate four year periods that showed significant variance, making it difficult to accurately forecast future fuel prices. When ignoring the variance, the model was not far off from reflecting the true growth of retail fuel prices.

The two periods of high variance were between 2005-2009, and 2013-2017. The initial period of variance had a sharp increase and decrease of fuel prices, more likely related to economic instability but this would need to be researched further. The second period of high variance had a sharp decline in fuel prices followed by a leveling off of the fuel price growth.

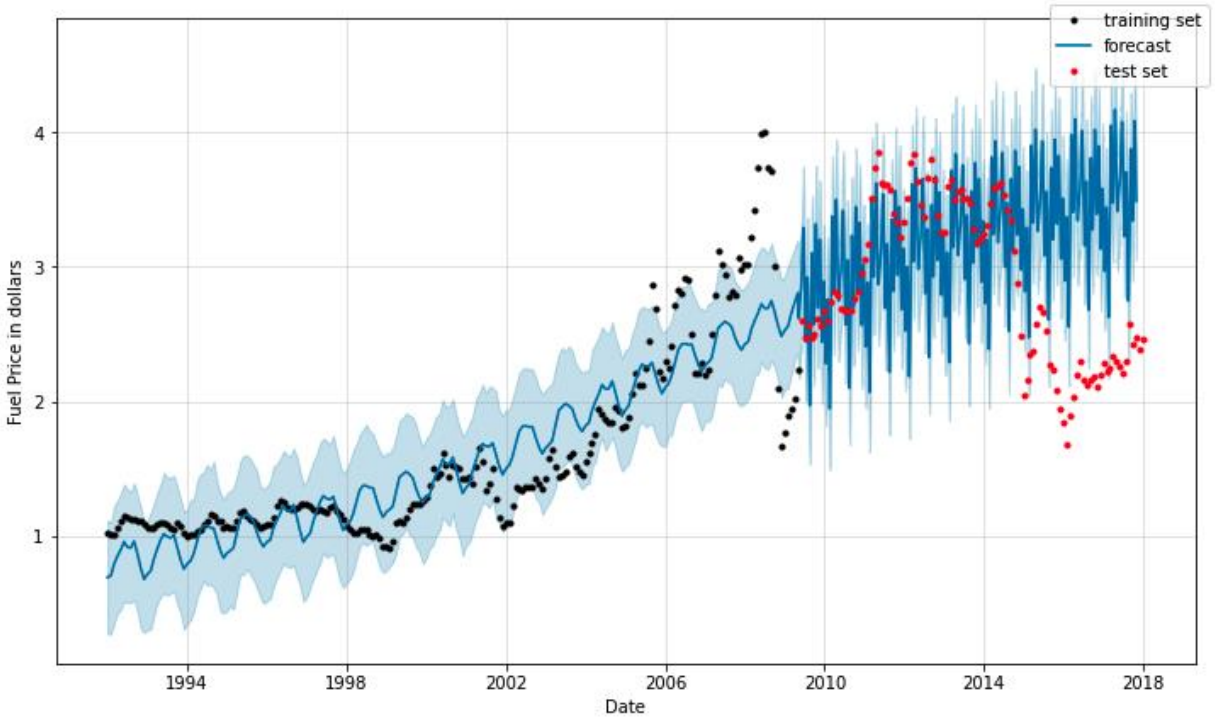
**Methods/Analysis:**

For the majority of the analysis, I leaned on Facebook Prophet's functionality and in-built modeling. To start, I used pandas read_excel() method to read through the U.S. Energy Review excel, and transformed it into a dataframe. I set a capacity on the resulting dataframe for later use in my modeling approach. I then split the data by a 70-30 split with 70% being the training set, and the remaining 30% as the test. I did this using scikit learns train_test_split() method from the model_selection sub library. After some further test, I was able to improve the model more by increasing the training set to 95% in order to account for both periods of high variance. I did give some consideration to cross validation, however, I was not certain on how to choose the amount of folds for this particular time series.

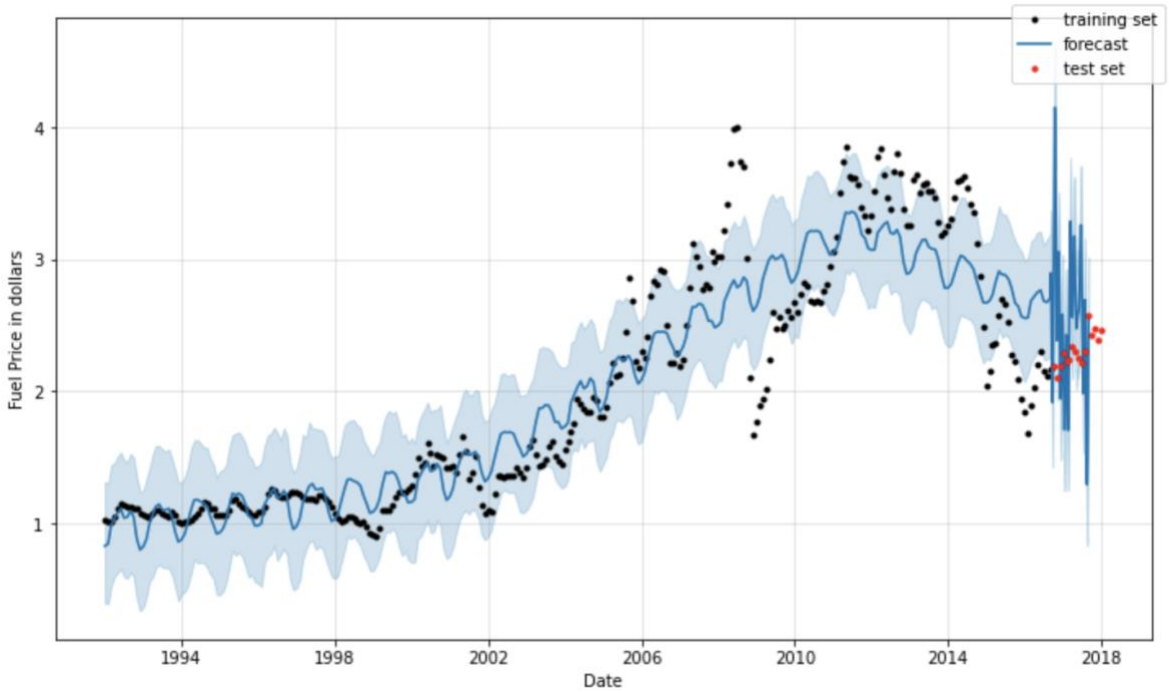
Using the training dataframe, I ran Prophets fit() method with the growth property set to logistic. I initially ran the fit() method using the default property of linear with the 70-30 split, but this resulted in a forecast that continuously increased to fuel prices that would not be considered reasonable. Referring back to the capacity I set, I made the assumption that gas prices would not exceed $4 in the given time range. With this assumption, and a growth model of logistic, I was able to better fit the test set of data, not including the significant variance of 2013-2017. I then decided to start pushing the training set percentage while running more tests in jupyter notebooks. After pushing the training set from 90% to 95% and switching the model growth back to linear, I noticed that the training set included both periods of high variance, resulting in a best fit model.

If I were to revisit this analysis and modeling, I would do more in depth research on how to address the two periods of high variance in the time series. I did expand the training set size to try and address this issue, however, I am concerned that this model may be overfitting if more data was introduced to the test set. Otherwise, the forecast model nearly predicts the fuel prices within the training and test set. Forecast model graphs, and instructions for re-running code can be found on next two pages.

70-30 Train versus Test Split




95-5 Train versus Test Split



**Instructions for Running Code/Notebook:**

To build the Docker image for the repository, use the driver.sh command below:
*Bash driver.sh build*

To create the corresponding container with the Jupyter notebook for analysis, run the following command:
*Bash driver.sh jupyter*

If you find that the Jupyter notebook command is opening a notebook that requires a password, you more likely have a Jupyter notebook running on port 8888 that has not been killed properly. Follow the link below to see commands for scanning for port 8888 processes, and how to kill processes with given PIDs.

https://stackoverflow.com/questions/3855127/find-and-kill-process-locking-port-3000-on-mac?page=1&tab=votes#tab-top