Forecasting Gasoline Prices with NLP Supplementation

Alex Walsh, Data Scientist - 2023/02/06

What is forecasting?

Forecasting is the science of trying to predict the future, usually used for quantities that are chaotic, pseudo-random, or otherwise just hard to predict

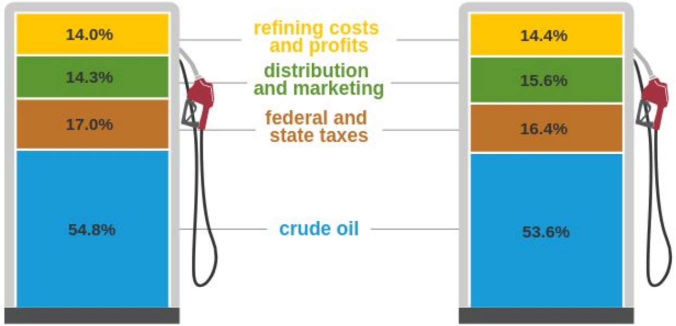
https://clipartix.com/wp-content/uploads/2016/05/Weather-clip-art-for-teachers-free-clipart-images.png

https://investdata.com.ng/wp-content/uploads/2017/01/TA_COURSES_STOCKS1.png

Price prediction

There are several techniques which exist that can make somewhat accurate predictions on prices

Recurrent Neural Networks (RNNs), ARIMA models, and linear forecasts can all be used to predict future values. However, all of these techniques fall short



eia Data source: U.S. Energy Information Administration, *Gasoline and Diesel Fuel Update*

Why do time series models struggle with predictions?

Missing data. This is where NLP techniques can help

NLP Supplementation...

- Can be used to infer data that is not numerically recorded anywhere

- Can help account for the “human” factor in the future of the data

...is very, very difficult

- Classifying text based on changes in a loosely-related quantity isn't the most hopeful endeavor
- A massive amount of text data would be needed to make this feasible
- My models were able to perform above baseline, but only by a narrow margin

How to source the data?

The Global Database of Events, Language, and Tone aggregates sources of events in the world and tries to classify who was involved, what happened, where it happened, etc.

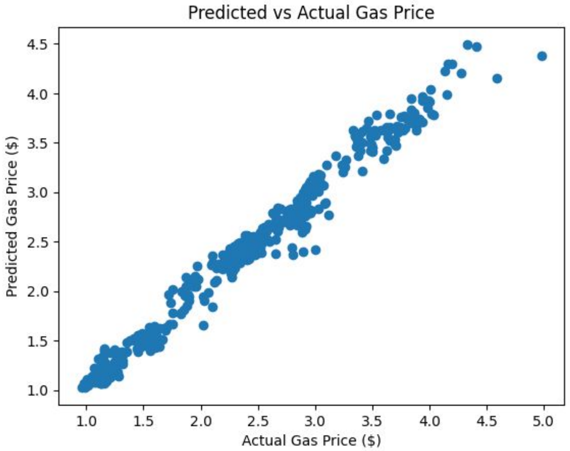
Using their events database, I aggregated several thousand news articles involving gasoline and/or oil

Back to numerics

Numeric data was sourced, including historical gas prices, various oil prices, oil reserves, and inflation

Gas price is strongly determined by these factors for a point in time

It is likely that future gas prices will prove to be predictable because of this



Feature engineering

An abundance of features was created to ensure as much information was captured as possible

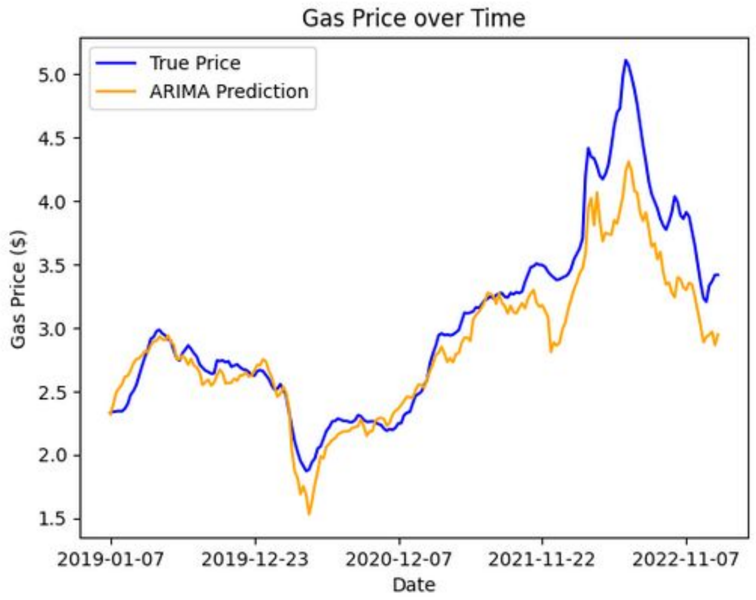
These features included lagged data, percent changes, Fourier components, price momentum, and rolling statistics among other things

These features were then fitted with Kernel PCA and reduced to five percent of their original size

Predictions: ARIMA

Autoregressive Integrated Moving Average (ARIMA) models are a statistical model which try to predict future values by weighting past values

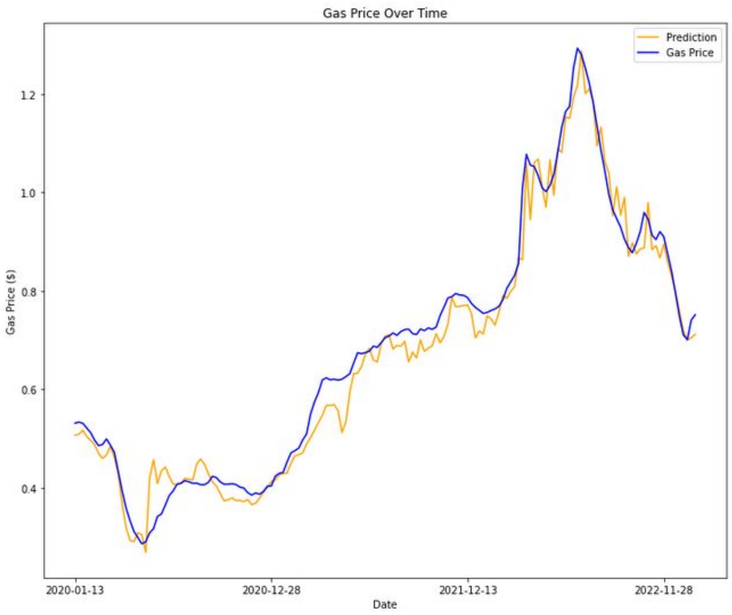
Performance of this model was decent, but it struggled to adapt to significant changes



Predictions: RNN

A second attempt at modeling used a recurrent neural network with ~13 million trainable parameters, consisting primarily of GRUs

This model followed closer to the general trend of the data than ARIMA did, but produced a very noisy output



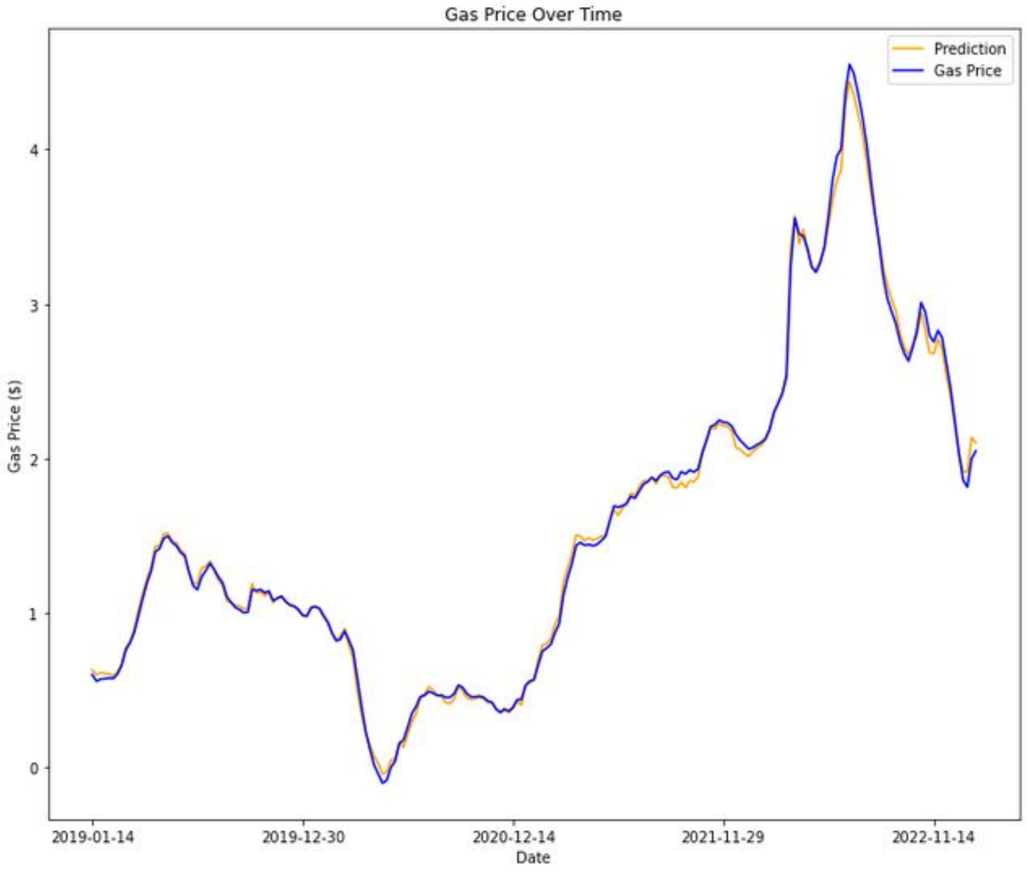
Putting it all together

- The RNN was trained on the full set of the numerical data, and was then used as one component of a larger model

- The larger model merged the GRU predictions with the ARIMA forecasts, plus some linear forecasts

- This is also where the NLP predictions were incorporated

The results



Mean Absolute Error: $0.035

RMSE: $0.048

Conclusions and Next Steps

- The modeling techniques I implemented proved effective for short term predictions
- Longer term predictions should be attempted with an autoregressive network
- The NLP components should be scrapped and rebuilt, likely using transfer learning, my attempts in this project barely scratched the surface of the potential benefit it brings

Summary

- I combined traditional forecasting techniques with NLP classification in an attempt to improve predictions

- No single technique worked especially well, but a combination of several in a deep model was effective

- The model was very accurate on the test set, with a MAE of $0.035