Colin Michael

DSC 630-T301 Predictive Analytics

Course Project: Milestone 1

I have confirmed that I will be doing my project on gasoline prices in the United States from 1995 to 2021. I gathered this dataset from Kaggle.com. It includes weekly gas prices for 13 different types of data.

I am fascinated by gas prices because of how greatly they impact America. Gas prices are often the subject of negative medium coverage for how rapidly they rise. Millions of Americans are forced to pay rising gas prices, which can place an extreme financial burden on people. Gas prices also impact shipping costs for businesses. This can cause significant increases in the total prices of goods.

My dataset from Kaggle is in csv format. I will import this data into Jupyter with the pandas.read\_csv function. The dataset includes 1,361 rows of data, one for each week from 1995 to 2021. Each row features 13 columns of gas prices. The data columns are gas prices for: all grades all formulations, all grades conventional, all grades reformulated, regular all formulations, regular conventional, regular reformulated, midgrade all formulations, midgrade conventional, and midgrade reformulated. This provides a comprehensive picture of average gas prices across different types.

Evaluating your data and determining your goal is the first step towards selecting a data science model type. My data set is numerical data over a period. My main goal is to use that data to predict future values. My first idea is to do a time series model. A time series is a series of data points ordered in time. In time series models, time is usually the independent variable, and the goal is to forecast the future (BuiltIn.com).

There are a couple of key factors to consider when running a time series model. The first is seasonality, which refers to periodic fluctuations in data. It will be interesting to see if gas prices are impacted by seasonality. My data set should do a nice job at revealing these trends.

Next is stationarity, which refers to a time series where statistical properties do not change over time. A data set is stationary if the mean and variance are constant (BuiltIn.com). Most time series models feature stationary data. I will perform a Dickey-Fuller statistical test on my data to determine if the time series is stationary.

I will begin my time series model by using an exponential smoothing test. This model calculates a moving average with the data and gives more importance to more recent data compared to older values. I believe that with inflation I will need to use this type of model, instead of a simple moving average model. I would also like to make a seasonal autoregressive integrated moving average model (SARIMA) time series. SARIMA applies seasonality to the following three components: autoregressive (a model that uses the dependent relationship between an observation and several lagged observations), integrated (the use of differencing raw observations to make a time series stationary), and moving average (using dependency between an observation and a residual error from a moving average model) (WisdomGeek.com).

I hope to use these models to better understand gas price fluctuations and predict future values. I believe that the SARIMA model will be more accurate than the simple exponential smoothing model because it considers seasonality. I believe there will be significant trends in data over the different seasons due to changes in demand over the seasons. For example, people typically travel more around the holiday season, so I’d expect gas prices to see peaks around the December-January months.

Using data ethically is arguably the most important part of data science. It is crucial that we as data scientists do not use data in harmful ways. One potential concern with my data model would be that it would reveal trends that would allow data companies to gouge their prices. Price gouging is when retailers rapidly increase the prices of goods. In May of 2022, the United States House of Representatives passed a bill to allow the Federal Trade Commission to investigate energy companies for price gouging (CNN.com). Price gouging inflicts great harm to consumers, and I will be careful to not structure my data model to encourage such malicious behavior.

My contingency plan if the exponential smoothing and SARIMA models does not work out would be to run a Prophet time series model. Prophet is an open-source software created by Facebook’s data science team and is available on R and Python (Facebook.com). It works extremely well with weekly data and is effective at revealing seasonal effects in data. The main appeal to Prophet is that it is considered easier to use and allows more clarity into seasonality in data.

I am very excited to dive into these models and the data. I have never run a time series model independently, so this should be a great learning experience. There are a lot of resources online and samples of code on time series models. I believe these will be a great resource for me as I run the exponential smoothing, SARIMA, and potentially the Prophet time series models.