****

**Predicting Retail Gasoline Prices at a Weekly Frequency in the United States**

**(Time Series Forecasting)**

**1- Problem Statement**

**Gasoline is a widely used motor fuel globally, with North America having the largest market share, followed by the Asia-Pacific region. The increasing demand for personal cars and small aircraft of private planes is expected to drive the growth of the gasoline market. The average global gasoline price is currently $1.32 per liter, $93 per barrel, with significant price disparities between countries due to varying gasoline taxes and subsidies. According to Data Bridge Market Research, the gasoline market was valued at USD 125,000 million in 2021, is expected to reach USD 140,811.6 by 2029, and is expected to grow significantly in the coming years.**

**Gasoline prices are a major concern for both consumers and businesses due to their impact on the economy. They have significant influence on nearly everything, from food to smartphones. Dorothy, Neufeld (2022) notes that oil prices have a considerable impact, accounting for up to**[**64%**](https://www.sciencedirect.com/science/article/pii/S0301421518308486)**of food price movements and have far reaching effect on the world economy, contributing to rising global inflation, slower global growth, and rising food insecurity and social unrest. Therefore, accurately predicting the fluctuations in gasoline prices and understanding the dynamics of pricing trends are essential for both individuals and organizations to make informed decisions.**

**In this report, we will provide a detailed analysis of historical gasoline price trends and use this information to make predictions about future fuel prices. The report will be useful for individuals and organizations who want to gain a better understanding of gasoline prices and how they are likely to change in the future. It will also provide valuable insights for policymakers and industry experts.**

**Research Question: Given the price history in last 23 years, what will be the optimal forecast value for gasoline prices for next 3 months starting from 18th of July 2022 to 8th of October 2022?**

**2- Dataset**

**The detailed datasets are provided by variety of resources, but for this study the dataset presented by California Energy Commission will be used. The California Energy Commission is the energy policy and planning agency that aims to influence policy makers to prepare for a future powered entirely by 100% clean energy. The dataset they provide contains weekly updates on gasoline prices in USA over the past 23 years, comprising 1228 weeks from January 4, 1999, to July 11, 2022. Only three updates are missing updates from the dataset: May 8, 2000, September 4, 2000, and August 27, 2001. The dataset includes two price values: Branded Fuel Prices (BFP) and Unbranded Fuel Prices (UFP), and it encompasses the following variables for both BFP and UFP):**

1. **Date**
2. **Branded Fuel - Distribution costs, and profits**
3. **Branded Fuel - Crude oil cost**
4. **Branded Fuel - Refinery costs and profits**
5. **Branded Fuel - State underground storage tank fee**
6. **Branded Fuel -State and local sales tax**
7. **Branded Fuel - State exercise tax**
8. **Branded Fuel - Federal exercise tax**
9. **Branded Fuel - Average retail prices**
10. **Unbranded Fuel - Distribution costs, and profits**
11. **Unbranded Fuel - Crude oil cost**
12. **Unbranded Fuel - Refinery costs and profits**
13. **Unbranded Fuel - State underground storage tank fee**
14. **Unbranded Fuel -State and local sales tax**
15. **Unbranded Fuel - State exercise tax**
16. **Unbranded Fuel - Federal exercise tax**
17. **Unbranded Fuel - Average retail prices**

**3- Methodology**

**Time Series is a time dependent method that analyses a collection of data points at constant time intervals. There are 3 components of Time Series: trend, seasonality, heterodaskisity (error). It is a very practical method to forecast a group of observations on a single entity over time. In many Time Series, there are seasonality trends, e.g., variations to a specific time frame. For instance, if you see the prices of stock value of skiing equipment fluctuates over time, you expect to find higher prices during Winter seasons. Time Series forecasting involves utilizing a statistical model to make predictions about future values of a given time-series by analyzing past results.**

**There are three major prediction tools that have been used in this time series analysis:**

**1- ARIMA forecasting algorithm relies solely on past values of a time series to forecast future values.**

**2- Exponential Smoothing forecasting technique predicts the next period value by utilizing past and current values involves data averaging, wherein the non-systematic components of individual observations are balanced out, and for short-term predictions, the exponential smoothing method is commonly employed.**

**3- Facebook Prophet is a robust forecasting tool that utilizes an additive model incorporating yearly, weekly, and daily seasonal patterns, along with holiday effects. It is particularly effective for time series data with significant seasonal trends and ample historical data. It can handle missing data and changes in trends, and outliers effectively.**

**4- Data Wrangling**

**The analysis begins with loading and exploring the dataset, which includes the weekly prices of gasoline in US from 1999 to 2022. The raw dataset consisted of 1228 rows and 17 columns. The size of the dataset was normal, and there was no need to reduce it. In the data preparation stage, the date column is converted to a datetime object and is set as the index of the dataset. Week, Month, and Year columns are created to investigate the trends in each.**

**3 missing values but no significant outliers are identified in the dataset. Since there are no sudden or drastic changes in the gasoline prices from one week to the next, and the previous week's prices are a good approximation for the missing values, the missing cells are filled in with the prices from the previous week.**

**Since the time series analysis will focus only on price prediction, the columns other than branded fuel - average retail prices and unbranded fuel - average retail prices are dropped from the dataset. Since these included exact same prices only one of them included in the data and the column is renamed as Price whose values then be converted from dollar to integer.**

**5- Exploratory Data Analysis (EDA)**

**Chart

Description automatically generated**

**When working with time series data, it is important to check if the data is stationary. After plotting the data to see whether the price varies with time, and using Augmented Dickey-Fuller Test, a stationarity is not identified in the data. This means that there is an increasing trend in the data and the series is a random walk. A transformation is applied to eliminate the trend from the data. After applying the difference and the log at the same time, the transformed series looked**

**stationary. This approach generated good results, with a ‘Test Statistic’ value of -1.100949e+01 and ‘Critical Value (1%)’ of -3.435735e+00 ensuring the stationarity in the dataset.**

**6- Algorithms and Machine Learning**

**This is a Univariate Time Series forecasting study and only the previous values of the time series are used to predict its future values. The data has been tested using three models: ARIMA, SARIMAX, Simple Exponential Smoothing, Holt-Winters Method, and FbProphet. Different metrics are used to evaluate the performance of the models, such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and** m**ean absolute percentage error (MAPE). MAPE is scale-independent metric, and it is easy interpret. That’s why it was the major metric used to select the winner model. It can also have heavier penalty for negative than positive errors, but this will not be a problem for this data because it does not have any negative value for the price.**

**According to the analysis of these three models, ARIMA has the lowest MAPE score and presents the best model accuracy. More details about each model are presented in the following.**

**ARIMA**

**First, an optimal ARIMA model is built and extended it to SARIMAX model. Then, the data is manually splitted as follows:**

**--Train data – January 4, 1999, and October 23, 2017 (First 18 years)**

**--Train + validation data - January 4, 1999, and October 21, 2019**

**--Test data - October 28, 2019, and November 7, 2022. (Last 5 years)**

**Graphical user interface, chart, line chart

Description automatically generated**

**The ARIMA model has three different components: autoregression, integration, and moving average and it is specified by three order parameters, namely p, d, and q, which refer to the order of the Auto Regressive (AR) term, the number of differencing required to make the time series stationary (I), and the order of the Moving Average (MA) term, respectively. These p, d, q values are essential in determining the accuracy of the ARIMA model. After determining the values of p, d and q, (2, 2, 2) ARIMA model is implemented to the data using the statsmodels library. The results for model performance are as the following:**

**MSE: 0.02 MAE: 0.13 RMSE: 0.16 MAPE: 0.1**

**Chart

Description automatically generated**

**Table

Description automatically generated**

**Based on the outcomes of the tests in the summary, the residuals are not correlated, and they are not normally distributed.**

**Exponential Smoothing – Holt Winters Method**

The idea behind Holt’s Winters method is to apply exponential smoothing to the seasonal components in addition to level and trend.

**Chart

Description automatically generated**

MSE: 1.50 MAE: 1.062 RMSE: 1.23 MAPE: 0.13

**Facebook Prophet**

**Facebook Prophet is a**n additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. IT is accurate, fast and fully automatic and it produces adjustable forecasts.

**Chart, histogram

Description automatically generated**

MSE: 0.62 MAE: 0.66 RMSE: 0.79 MAPE: 0.14

**7- Conclusion**

This time series analysis project was designed not only to better understand how gasoline price changes over time, but specifically to utilize machine learning tools to predict the future price values. According to the analysis of these three models, ARIMA has the lowest MAPE score and presents the best model accuracy. Therefore, we selected ARIMA as the best model to predict gasoline prices.

**8- References**

**Data source: California Energy Commission. Accessed at** <https://www.energy.ca.gov/media/5893>

Global [Gasoline](https://www.databridgemarketresearch.com/reports/global-gasoline-gensets-market) Market, By Type (Regular Gasoline, Special Gasoline), Application (Motorcycle, Automobile, Others) - Industry Trends and Forecast to 2029. Accessed at: <https://www.databridgemarketresearch.com/reports/global-gasoline-market>

**Dorothy Neufeld (2022). The inflation factor: how rising food and energy prices impact the economy. Accessed at** <https://www.visualcapitalist.com/inflation-rising-food-energy-prices-impact-economy/>

# **Prashant Banerjee (2021).** Tutorial: Time Series Forecasting with Prophet. **Accessed at:**

# <https://www.kaggle.com/code/prashant111/tutorial-time-series-forecasting-with-prophet>