## Contents

# Forecasting Inflation using Alternative Data and Machine Learning

Dimitry Budarin

October 15, 2022

**Abstract**

The aim of this study was to use the latest in machine learning techniques alongside an alternative data source in order to assess their effects on the ability to forecast inflation. Using Random Forest and Automatic Identification System (AIS) data, which contains second-by-second ship location information from around the USA, the results of the study show a comprehensive improvement over the benchmark models. The study also aimed to use the latest explainable AI methods to dwell deeper into understanding of the decision-making process of the model. The explainable AI was indeed able show the logic behind the model with mixed results. Finally, it is important to note that the data used, only covered a period of seven years during which, the COVID pandemic had created unprecedented economic conditions. It is therefore necessary to carry out further research ideally with bigger dataset, covering longer time periods, in order to determine feasibility of the forecasting model across different horizons.

## 1 Introduction

Inflation is one the most important macroeconomic phenomena that requires constant and careful control. If it is left unchecked, runaway inflation will create economic instabilities whenever it persists. Too much, and inflation will cause erosion of currency's purchasing power leading to hardships for the population struggling to pay their everyday bills. Too little, and this could indicate an inefficient monetary policy that is slowing economic growth eventually leading to a recession. Maintaining this balance is the responsibility of the central banks. To achieve this, the central banks try to accurately forecast inflation in order to take the most effective decisions. These actions directly impact all stakeholders of the society, from workers and businesses, to governments and central banks themselves. Accuracy of such inflation forecasts is therefore absolutely crucial.

In most developed nations, the inflation is targeted to remain at approximately 2%. The central banks use various tools, most notably short-term interest rates, to maintain this target. For the last 30 years, inflation has been stable, staying close to the target rate in the developed world. However, since the pandemic of 2020, the inflation has soared to 8.6% in the USA as of May 2022, and even further in some European countries – 40 year highs. Presidents of the central banks like Jerome Powell of the FED and Christine Lagarde of the ECB, seem to agree that this is the result of numerous factors coming together [27]. These factors range from being demand-pull, where the savings of the population and the stimulus provided by the governments created too much demand, to supply-push, where the supply chain bottlenecks caused shortages of goods.

Fortunately, there are numerous econometric methods and techniques, which could be used to forecast inflation. Recent developments in Machine Learning (ML) have demonstrated the viability and effectiveness of this technique with promising results for predicting inflation forecasting. Advanced models using Random Forest and XGBoost have been shown to outperform classic models that use VAR or AR regressions [17]. In 2020, developments in ML model explainability by researchers, Bowen and Ungar, provided further confidence into the results of such models. With their techniques, the decision-making process of these models could be rationalised thus addressing the long-standing notion of them being unexplainable "black boxes" [3].

Another area, which recently has seen significant developments, is "the Big Data". In the past decade, data had started to be collected in great quantities from the most unusual of sources to solve problems in the most creative ways. This ranged from texts being web scrapped from tweets that were then processed into quantifiable sentiments, to monitoring traffic using satellite imagery in order to estimate economic activities of businesses, to tracking population movements using mobile phone location data. This type of data has become known more generally as the "Alternative Data", and it is most commonly characterised by its hard-to-process nature requiring a lot of cleaning from its raw format [12]. As the world becomes more and more digital, this type of information is only more likely to become more abundant. With appropriate cleaning techniques, this data represents an opportunity to obtain a vast array of insights that could be used to improve forecasting, especially on the short-term basis as demonstrated by Dessaint et al [13].

The ML applicability has evolved in parallel "with the Big Data". To an extent, it could be argued that "Big Data" is partly responsible for the successes that ML models have achieved, as these non-linear models are ideal to process vast volumes of data across many variables.

One of these "Alternative Data" sources is the Automatic Identification System (AIS) used to track shipping traffic throughout the world. This system has become a standard on big vessels since the early 2000s. Nowadays, almost every ship is required to have a transponder that is connected to the AIS. The AIS database is updated with the location and the speed of the vessel, amongst other data with frequency of up to every two seconds. This information is then used by ships and shipping companies to control traffic and improve their situational awareness. With all the advances in data science, the AIS data has started to show practical uses in economic models with strong potential to even be useful in forecasting macroeconomic variables.

Using the latest Machine Learning methods, this paper aims to use the US AIS data in combination with other market and fundamental indicators to determine how much of an improvement, if any, can be obtained to forecast inflation. With the latest developments in explainable Machine Learning, this paper will also set out to clearly demonstrate how much of an impact the "alternative data" has on the model and its predictive capabilities, especially in the more volatile environment of the last two years. If successful, this could pave a way to use such data to forecast inflation in other countries where conventional data is hard to obtain or is unavailable.

## 2 Literature Review

Research into forecasting inflation has been ongoing for many decades, initially starting off with simple regression models and then recently moving to more complex ML methods with some starting to use forms of alternative data.

In 1990, Hafer and Hein compared the performance of the inflation forecasting models when using short-term interest rates with univariate autoregressive models across six countries [21]. Interestingly, the study showed that by themselves, the autoregressive (AR) models were more accurate than when the models only used a short-term interest rate, however the models that combined the two, showed a slight outperformance in five out of six countries. These early findings suggested two things: AR models are already rather efficient at forecasting inflation and at the same time that they could still be improved upon by introducing other variables into the model.

This was further demonstrated with more sophisticated models, such as those proposed by T. Clark (2011) using BVAR [10] and D'Agostino's et al (2009) VAR model [11] that included exogenous variables such as unemployment rate, interest rates, and GDP deflator. They observed more accurate forecasting compared to the simple autoregressive models, outperforming them by 30%. These results also suggested that the advances in econometric techniques as well as new data have allowed for better macroeconomic forecasting.

Greenspan (1992), Bernanke and Woodford (1997), and Shiller (1996) have all previously used market data for inflation forecasting, mainly by observing the evolution of Treasury Yields on the Treasury Inflation Protected Securities (TIPS) as well as Inflation Swaps derivatives (Shiller, 2009) [19] [2] [6] [7]. They argued that these financial products could be used as inflation expectations gauges as priced in by the markets and are therefore a potential indicator of the direction of the inflation. According to Fama (1960) and reaffirmed by Malkiel (1989) [16] [23], this is a reasonable argument and is in line with the "Efficient Market Hypothesis" that states: "asset prices reflect all available information" – inflation included.

Others, such as Atkeson and Ohanian in 2001, tried to see if the well-known relationship between the inflation and employment in the Philips Curve could be used determine the inflation trend [1]. The research concluded that the use of Philips Curves in their models showed no forecasting advantage over the naïve models, which assumed the best predictor of inflation for the next year is the inflation of previous years, in other words the conventional AR models.

From the literature, it became quite apparent that the performance of these models varies significantly depending on the testing period that is chosen for the evaluation. This is largely because inflation has varied significantly during different periods of the last century. As Stock and Watson (2007) point out, the period of "Great Moderation", 1985 – 2007, which represented a period of low inflation volatility is much more favourable to the autoregressive models. The Root Mean Squared Error (RMSE), the metric most commonly used to measure the performance of these models, is significantly lower in AR models giving an impression of greater outperformance [29]. Faust et al (2011), suggest that in order to evaluate the performance of the model accurately, the testing period should have periods of higher inflationary volatility to ensure accuracy across different economic environments [17].

By 2011, there were numerous models and techniques proposed by researchers to forecast inflation. To summarise and standardise these findings, Faust and Wright carried out a review and compared 17 different forecasting methods using macroeconomic data. They used data prior to 1985 to train, and between 1985 and 2011 to test the models [17]. The post 2007 period was intentionally included in the testing to address the "Great Moderation" problem described earlier. Post 2007, inflation became more volatile after the Financial Crisis of 2008, thereby providing a more representative testing period. In their models, they focused on predicting the inflation based on the GDP deflator, Personal Consumption Expenditure (PCE), Consumer Price Index (CPI), and Core Consumer Price Index (CCPI). The methods used for the models included: AR, Philips Curve, Random Walk (RW), unobserved stochastic volatility model (UCSV), VAR, equal weighted averaging (EWA), Bayesian model averaging (BMA), dynamic stochastic general equilibrium (DSGE), and finally, the study also included three forecasts made by third parties (Blue Chip, Survey of Professional Forecasters (SPF), and FED's Greenbook). These third parties utilised private economic data that was passed through a "subjective filter", in other words through experts who could modify the results according to their own beliefs. The best performing model still remained the AR (if not include the three third party forecasts that were better). This was surprising, considering that the testing period was extended to include more volatile inflationary dates that should not favor the AR models. The better performance of the third parties was attributed to professional expertise as well as the "private" data access that these institutions would usually have. This "private" data is usually based on recent events such as natural disasters, wars, or political decisions, which would not necessarily be captured by the all of the variables in the models. For nowcasting, once again, third parties performed 20% more accurately than the best AR model, concluding that the results really depend on data availability and its appropriate use.

On a more technical level and something that should be taken into consideration for future research, Faust and Wright (2011) noted that non-stationary models performed better than their stationary counterparts [17]. They explain this by the fact that up until 2011, the CPI has been persistently lower than the historic average leading to overestimation by the models. Additionally, the UCSV models performed best with smoothed CPI data, as also noted by Stock and Watson (2010) [30].

Since the review by Faust et al in 2011, "a new breed" of predictive methods have started to be commonly used in forecasting inflation, these most notably consist of Machine Learning (ML) techniques [17]. Medeiros et al (2021) carried out a comparison study between different ML methods as well as the classic regression models. Using the "FED macroeconomic database" with 127 variables over the last 60 years, the authors concluded that ML models, above all - Random Forest, were able to outperform the AR models by up to 30% [24].

Despite their outperformance, an area where ML models have been heavily criticised has been explainability. Up until recently, these models have been viewed as "black boxes", however developments by the "SHAP" team have demonstrated that the ML decision making process can be quantified (Bowen and Ungar, 2020) [3]. They have shown that by using the game theory approach, a shapley value can be calculated for each variable in order to determine its importance and its weight on the final output. Others have developed partial dependence plot methods to measure the sensitivity of changes in variables on the outputs to provide further explainability to ML models (Molnar et al, 2021) [25]. As it stands, these techniques appear to have been completely overlooked by the current ML research into inflation forecasting and should be investigated further.

Finally, the uses of "Big Data" and "Alternative Data" in macroeconomic forecasting have also been reviewed. It was found that some research had started to shift away from using traditional macroeconomic data and towards more “alternative” sources in order to predict inflation more accurately and with higher frequency. In 2015, A. Cavallo published a study nowcasting inflation figures based on “alternative data”, where he was able to web scrap product pricing data from supermarkets’ websites on daily basis from 2008 until 2015 [9]. From this data he was able to create a basket of goods, which could proxy the CPI, an in a sense nowcasting it in real time. Cavallo concludes that frequency of price changes is a good correlation metric with the inflation.

Insights were also captured from “alternative sources” by Picault et al (2022) by estimating European inflation expectation sentiment from newspaper articles based on ECB communication [26]. The research found that the lagged calculated media sentiment had a significant effect on the inflation expectation during the next period.

Fornaro and Loumaranta (2018) combined the use of ML methods and alternative data. In this case, they used Finish traffic data to nowcast Finnish economic activity [18]. First, they evaluated the performance of the machine learning models using standard economic indicators, followed by models that included the high frequency traffic data. Both models performed as well as the official estimates and were deemed to be sufficient for nowcasting.

Studies using different shipping data have also been performed. Shipping costs data were used by Carrier-Swallow et al to demonstrated its effectiveness in estimating future inflation [8]. Furthermore, recent research by LaBelle and Santacreau gives an indication of how AIS data could be used in forecasting inflation [22]. The study concluded that in 2022, supply chain bottlenecks have contributed up to 2% towards the current inflation by creating an environment where the demand far outstrips the supply. They identify the unprecedented post-pandemic reaction by the world as one of the main culprits for this phenomenon. The initial shock of the pandemic caused businesses to cancel orders and renegotiate their supply chains. However, the unexpectedly high demand for physical goods during the lockdowns created supply chain bottlenecks. In a country such as the USA, 80% of all goods usually arrive by sea, making ports some of the prime candidates for the locations of these bottlenecks. If this is indeed true, and if it is possible to monitor and accurately estimate the changes in shipping traffic characteristics, it could potentially also give insights into the evolution of inflation.

What can be concluded from this literature review is that there is a clear evolution of inflation forecasting models throughout history. On one hand, the methods had evolved from using AR, which still outperform many “sophisticated” models, to Machine Learning that uses Random Forest or XGBoost. On the other hand, the data, which is being used has also evolved, from only using short-term interest rates to using 127 variables, daily market data, and finally the alternative data. This paper aims to continue in this direction, by exploring further into use of Alternative Data and ML. Using AIS shipping data in combination with Random Forest ML methods. Finally, it has also been noted that ML model explainability has been remarkably missing from all of the ML inflation papers that have been reviewed. This research will add to the literature by aiming to forecast inflation using AIS data, which has not been attempted before, and finally it also aims to clearly show the logic of the ML models by showing each feature importance and how each variable contributes to the final output to better understand the decision making process of the ML models.

# 3 Data

Data should be considered as the most important part of the model. Clean, accurate, and appropriately transformed data ensures accuracy of the results. In this research, three data sources were used to create one dataset, which was used by the model. This study created two sets of dataset: one consisting of monthly and one with daily variables. The reasoning behind using two datasets is to address different frequencies of variables present in the data. Shipping and Market data is available at daily frequency, however the target variable is only available as a monthly figure. Ideally, the daily indicators would be converted to monthly, however this dramatically decreases the number of observations. For ML models, having high number of observations to train on, is rather crucial for an accurate model, therefore the target variables was interpolated to daily frequency. Models were ran on both, and two sets of results are presented.

## 3.1 Automatic Identification System (AIS) and World Port Index (WPI) Data

The AIS data, tracking locations of the ships, is collected and updated by the US Coast Guard Navigation Centre. Data itself was downloaded in its raw format from https://marinecadastre.gov/ais/. The data was available for the period of 2015-01-01 to 2022-03-31 at second frequency. Due to the sheer quantity of information and its raw format, downloading it, was a challenge in itself. Each daily file had to be downloaded separately, overall 2,280 individual downloads were made totalling a combined size of over 456GB. On average, each file had approximately 7,000,000 observations representing locations of around 14,000 ships around the coasts of US on a specific day. An example of the raw data is presented in Appendix A and visualisation of ship locations is presented in Figure 1 with a close up on the New York Port represented in Figure 2. Figures were visualised using "AccessAIS" software.

In addition to the location of the ship, each file contained the following metadata about the ship:

- MMSI (Maritime Mobile Service Identity) – Unique ID given to every vessel.
- Base Time Date – Time at which the transmission was received.
- Latitude/Longitude – position coordinates of the vessel
- Speed over Ground (SOG) – traveling speed of the vessel
- Heading – travelling direction of the vessel
- Vessel Name
- Vessel Type – code assigned to the vessel depending on its function (see Appendix 1)
- Status – code associated with vessel's current activity (see Appendix 2)
- Cargo – code associated with the type of cargo transported by the vessel
- Length/Width – vessel's dimension

## 3.2 AIS Data Cleaning

The overall goal of the Data Cleaning step was to identify the number of ships waiting at ports and analyse their waiting times. This was calculated by counting the number of ships at ports on each day

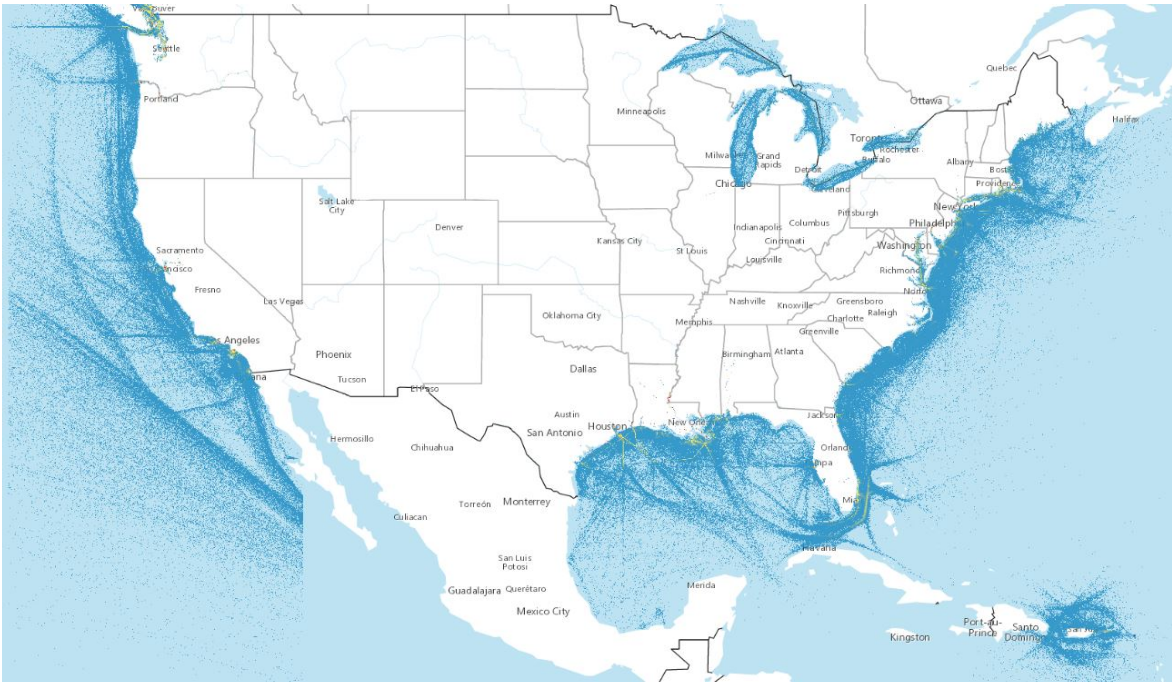


Figure 1: Ship locations around the coasts of USA on 2020-12-28 (dark blue spots represent the positions of the ships)

and then by calculating time spent there. Due to computing limitations, it was not feasible to process all the data at once. Therefore, it was necessary to reduce the size of each individual daily file before adding the data to the cleaned database.

Cleaning commenced by removing ships that were not considered useful to the study, in other words, ships, which were not waiting in the queue to the ports. Using the World Port Index (WPI) database, only ships within a close proximity to the US ports were selected. WPI contains information about the location and physical characteristics of ports around the world (example of raw WPI data presented in Appendix A). The database is updated by Maritime Security of National Geospatial Intelligence Agency and was downloaded from https://data.humdata.org/dataset/world-port-index.

Both, the AIS and WPI, contain latitudinal and longitudinal information of the ships and the ports respectively. By matching these coordinates, it was possible to filter the ships which were potentially waiting to be unloaded. However, as per Bloomberg (2021), it was noted that some ships were anchored up to 80 nautical miles away from the port waiting for their turn to unload. Therefore, a radius of approximately 40 miles was taken around the port's location as a zone of potentially interesting vessels. However, an issue was noted when there were multiple ports within 40 miles of each other that created duplicate results. Naturally, to maintain data consistency, one of the duplicate results was removed, nevertheless this meant that it was not possible to correctly assign this ship to a specific port. Therefore it has to be noted that this is not the most accurate method and could be improved upon, however on a national scale this was not considered to be an issue. Nevertheless, individual port waiting times and ship counts were still collected, however ports that had their areas overlap were excluded.

To filter the data even further, "Vessel Type", "SOG" (speed over ground), and "Status" parameters

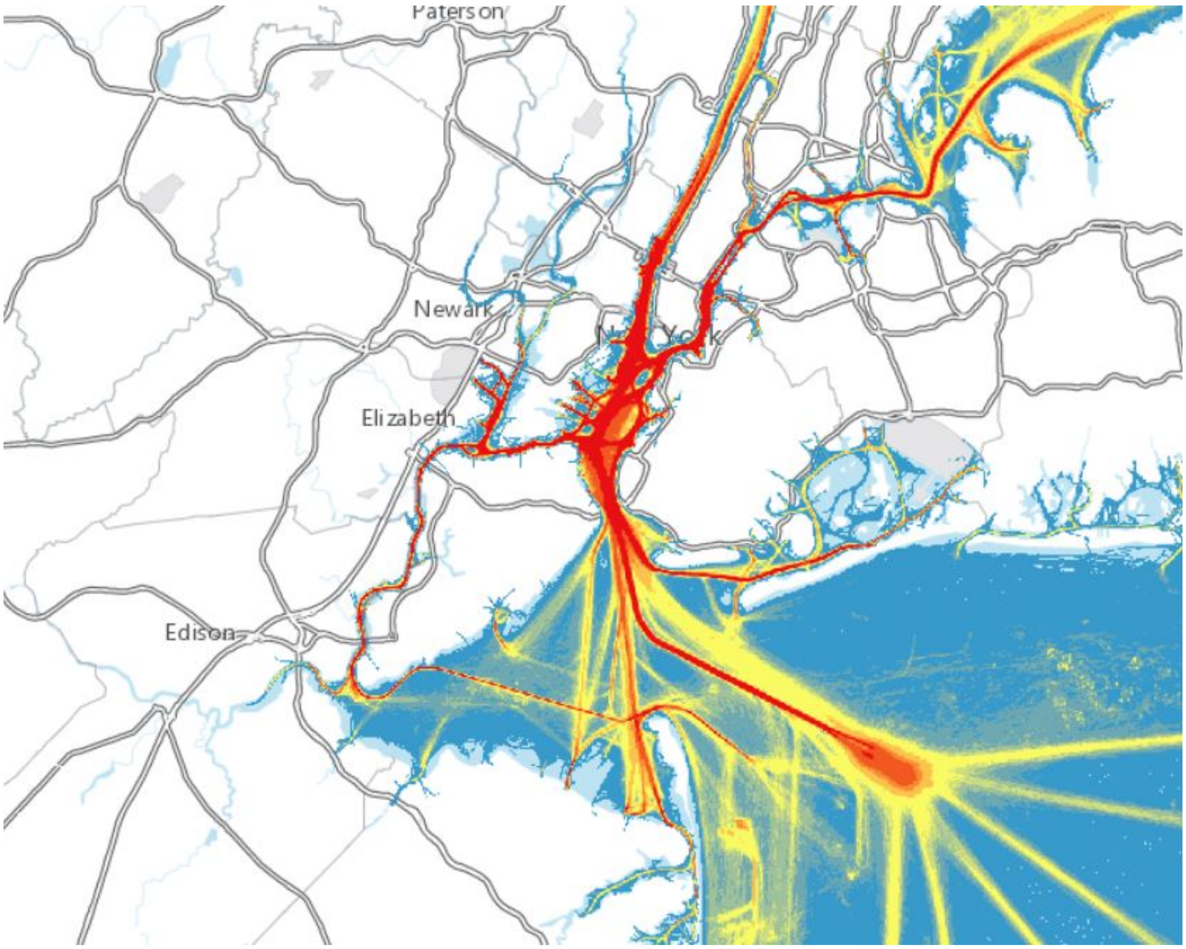


Figure 2: Close up of New York shipping traffic on 2020-12-28.

were also specified in the cleaning stage. Firstly, as the study focused on supply of goods as relevant to the inflation, only tankers and cargo ships were specified as a "Vessel Type". Secondly, out of these cargo and tanker vessels, only those with status as "Anchored", "Moored", or "Undefined" were chosen. This was assumed to be an indication that the ship was waiting to be unloaded. Vessels that were "anchored" or "moored" for more than three months were considered as non-functional and removed. The "Undefined" vessels were filtered based on their speed. If the speed was less than 1 mile per hour, the vessel was considered to be waiting to be unloaded. The Diagram of the Cleaning Process is presented in Figure 3. Finally, the data was down sampled from frequency of seconds, to daily. Overall, this reduced the daily dataset from approximately 7,000,000 observations to approximately 20,000.

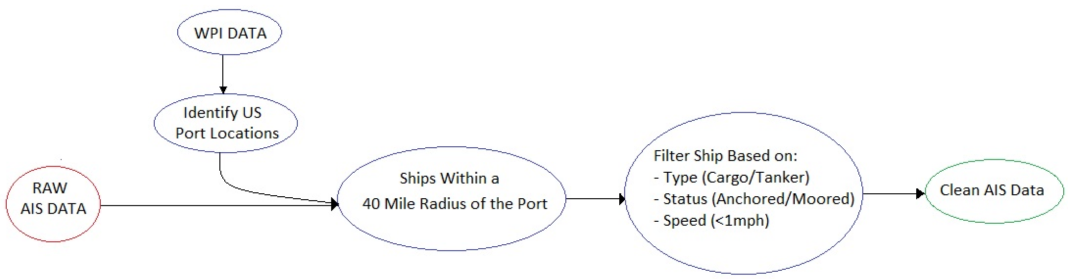


Figure 3: AIS Data Cleaning Procedure.

Once all the relevant vessels were identified and combined into one dataset, four sets of indicators were calculated at daily frequency (Table 1) in addition to individual top 10 "busiest" ports as identified by the number of daily ships from the data (Table 2). Moving averages of these indicators were calculated to smooth out the data and remove outliers as well as the unusable noise (Summary Statistics of the calculated shipping variables are presented in Appendices B and C, and the raw vs moving average graphic is included in Appendix D):

| New Global Variable Calculated | Calculation description | Figure |
|---|---|---|
| Daily Average Ship Count in US ports | Moving Average of 1 years | (Figure 4) |
| Daily Average Ship Count in top 40 US ports | Moving Average of 1 years | (Figure 4) |
| Average waiting time in US ports | Moving Average of 6 months | (Figure 5) |
| Average waiting time in top 40 US ports | Moving Average of 6 months | (Figure 5) |

Table 1: Global indicators derived from the AIS data.

| Port | US Trade (bil) (% of GDP) – 2013 Figures |
|---|---|
| Port Everglades | 21.7 (0.13%) |
| Bayonne (NY and NJ) | 123.3 (0.73%) |
| Gretna (Louisiana) | 238.5 (1.42%) |
| GALVESTON | 229.24 (1.37%) |
| Long Beach | 84.5 (0.5%) |
| Alameda | 19.3 (0.12%) |
| Baytown | 229.24* (1.37%) |
| Savannah | 31.9 (0.19%) |
| Miami | 7.13 (0.04%) |
| Norsworthy | NA - No Figures available |

Table 2: Top 10 Busiest Ports Identified from the Data. Source for the Trade Data: American Association of Port Authorities.

Descriptive Statistics are presented in Tables 3.

| | Mean Count | Mean Top40 | Port Time |
|---|---|---|---|
| count | 2635.00 | 2635.00 | 2635.00 |
| mean | 235.89 | 169.42 | 43.45 |
| std | 26.79 | 20.92 | 4.51 |
| min | 175.09 | 125.02 | 34.80 |
| 25% | 217.32 | 153.82 | 39.96 |
| 50% | 236.32 | 168.29 | 42.83 |
| 75% | 252.48 | 182.04 | 45.79 |
| max | 311.03 | 225.45 | 57.26 |

Table 3: Descriptive Statistics of Cleaned AIS data - Mean Ship Count, Mean Ship Count in the Top 40 Ports, and Average Ship waiting time.

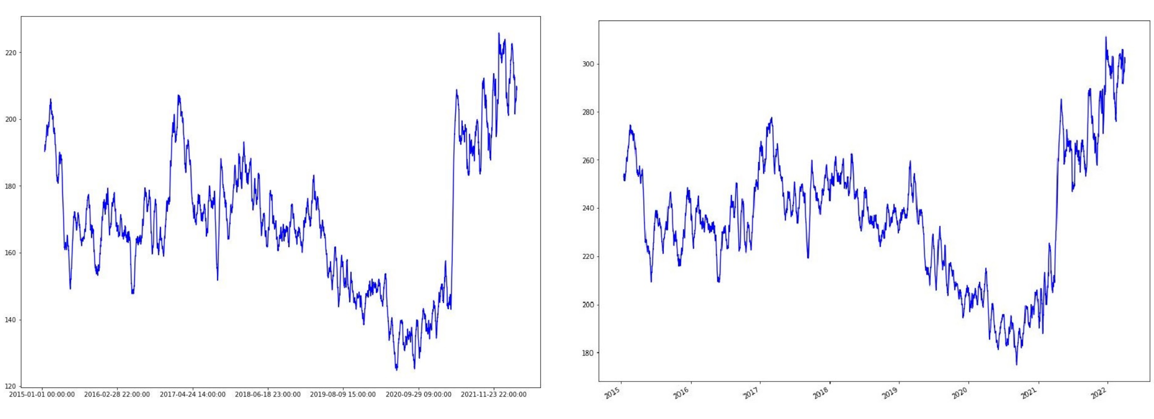


Figure 4: Daily Average Ship Count in top 40 US ports and in all the US Ports.

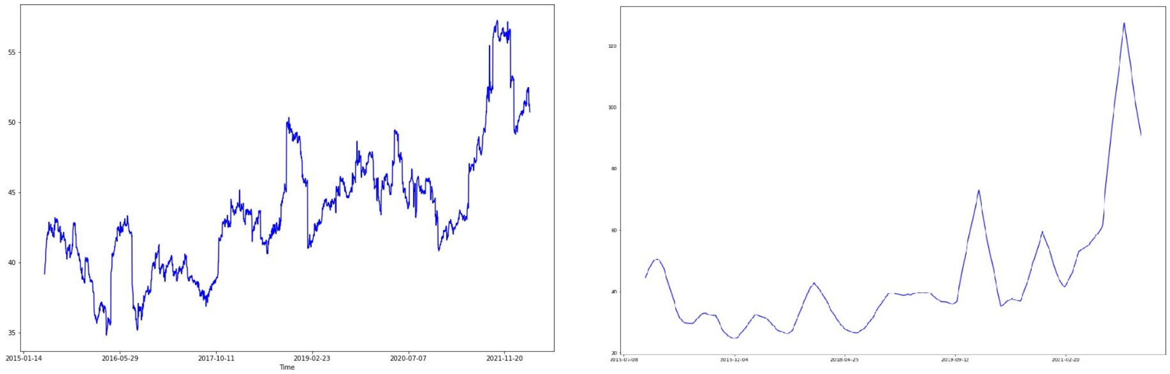


Figure 5: Average Ship waiting time in US ports and Average Ship waiting time In Alameda (SF) Port (example of an individual US port).

### 3.3 Bloomberg Data

In addition to the shipping traffic data, the models was also complemented by the market and economic data downloaded from Bloomberg. The period was limited to the same period that the AIS data was available. The database consisted of 13 market indicators are presented in Table 4 and summary statistics are presented in the Appendix B and C.

| **Variable** | **Description** |
|---|---|
| TIPS1 | Treasury Inflation Protected Security - 1 year yield |
| TIPS5 | Treasury Inflation Protected Security - 5 year yield |
| TIPS10 | Treasury Inflation Protected Security - 10 year yield |
| Oil Price | Brent Oil Price |
| ISMM | ISM Manufacturing Sentiment Index (PMI) |
| NFPTOT | Non Farm Payroll Total (New employment) |
| Gold/Copper | Ratio gold to copper |
| Inflation Swap 5Y5Y | Inflation Swap Rate |
| BED | Baltic Dry Index (Cost of shipping goods) |
| SP500 Performance | 30 Moving Average SP500 Returns |
| SLOPE 2-10 | 2 year - 10 year yield spread |
| M2 Growth | M2 money supply evolution |
| CRB | Commodity Index |

Table 4: Market and Fundamental variables used in the model (Bloomberg).

Monthly variables are primarily represented by sentiment indicators obtained from ISM surveys and include: PMI (ISMM) and Employment. The target variable, CPI, is also part of this dataset. Daily variables mainly consist of market data and range from Inflation Protected Treasury yields, Performance of US Equities, and Prices of Commodities. Variables were selected as based on previous research identified in the literature review. Daily data was obtained only for the business days. To create a daily dataset, monthly variables, including CPI, were interpolated using the Savitzky-Golay filter (Schafer, 2011) [28].

Data was tested for stationarity with AD Fuller test and transformed if necessary. CPI was shown to be non-stationary, therefore first differences were taken. As previously noted by Faust and Wright (2011) [17], stationarity does not seem to have as much of a significant effect in ML models as it does in traditional regression models, therefore models were run with both stationary and non-stationary variables. Both, stationary and non-stationary time series of CPI are presented in Figures 6 and 7, summary statistics are found in Table 5.

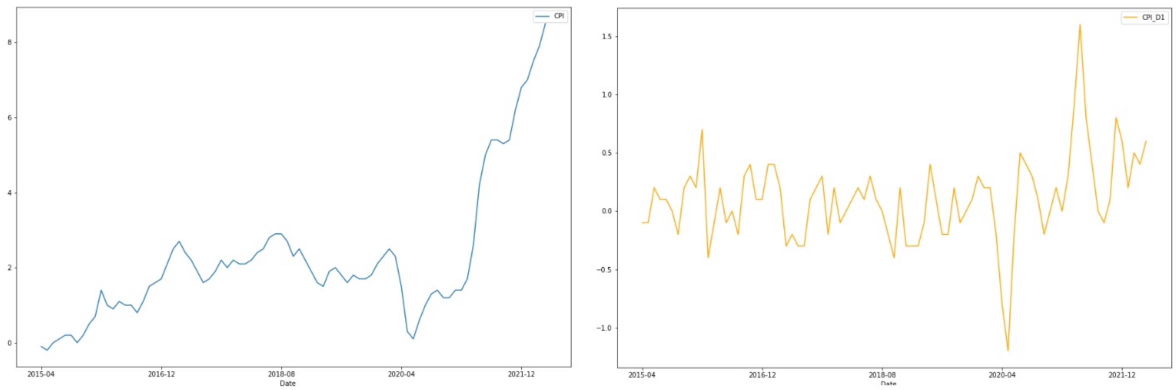


Figure 6: Monthly CPI (Non-Stationary and Stationary) - 2005-2022.

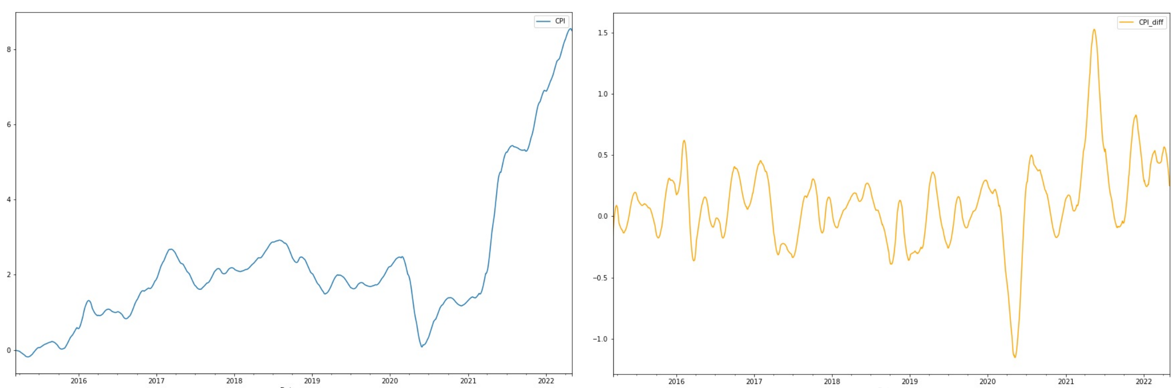


Figure 7: Daily CPI (Non-Stationary and Stationary) - 2005-2022).

| | **CPI_monthly** | **CPI_Diff_monthly** | **CPI_daily** | **CPI_Diff_daily** |
|---|---|---|---|---|
| count | 85.00 | 85.00 | 1865.00 | 1865.00 |
| mean | 2.22 | 0.10 | 2.21 | 0.10 |
| std | 1.86 | 0.37 | 1.86 | 0.35 |
| min | -0.20 | -1.20 | -0.19 | -1.15 |
| 25% | 1.20 | -0.10 | 1.15 | -0.09 |
| 50% | 1.80 | 0.10 | 1.81 | 0.09 |
| 75% | 2.50 | 0.30 | 2.45 | 0.27 |
| max | 8.50 | 1.60 | 8.55 | 1.53 |

Table 5: Descriptive Statistics of Daily and Monthly CPI figures.

# 4 Methodology

## 4.1 Counterfactual Inflation Impact Estimation

Prior to trying to perform inflation forecasting, a counterfactual study was performed to examine exactly how much of an impact, shipping bottlenecks had on the evolution of the inflation in the US. To estimate, a simple OLS regression was ran using the CPI as an endogenous variable and the lagged CPI and the mean number of ships in the top 40 ports as the exogenous variables.

The counterfactual started from March 2020, as this represented the start of the pandemic, the start of significant disruptions to the maritime traffic around the US, and the start of the rise in inflation. An average number of ships at the US ports was calculated prior to March 2020, which was 158 ships. This figure was taken as a "benchmark number" and was assumed to be the number of ships that US ports are able to process efficiently enough to prevent bottlenecks. After March 2020, differences were taken to the "benchmark number" to see the impact to the number of ships caused by the pandemic. These differences were then multiplied by the coefficients estimated from the OLS regression to estimate the impact caused by shipping disruptions to the inflation.

## 4.2 Benchmark Forecasting Models

### 4.2.1 Autoregression Models

Autoregressive models were chosen as the benchmarks for this study. Previous literature had demonstrated that these model consistently outperformed majority of other models with relatively good results (Faust and Wright, 2011) [17]. AR regression equation is presented below, where c is the constant, B is the parameter and e is the error term.

$$X_t = c + \sum B^i X_{t-1} + \varepsilon_t$$

Two AR benchmarks were set, one for the monthly and one for the daily frequency data. First differences of the CPI were taken as the variable for the monthly data, and 22-day (business days) difference for the daily model, simulating a forecast of one month ahead. Partial Autocorrelation Plots indicated that a lag of two was significant for the monthly and a lag of ten for the daily data (Appendix F). The parameters were then estimated using OLS. Testing period was set for the period of 2021-06 to 2022-05, representing 10% of all observations. Performance was then measured by calculating the Mean Squared Error (MSE). Lower MSE indicated better performance.

### 4.2.2 Vector Autoregression (VAR) Models

Vector Autoregression (VAR) models also served as a multivariate benchmark, with the lagged endogenous variable. These models were previously shown to yield promising results (D'Agostino's et al, 2009) and were therefore deemed as a useful benchmark for evaluating the performance of Machine Learning methods for forecasting inflation [11].

VAR models capture the inter-dependency between time-series based on the lags of all the variables in the data as per the equation below:

$$\begin{bmatrix} y_1, t \\ y_2, t \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \phi_{11}\phi_{12}.. \\ \phi_{21}\phi_{22}.. \end{bmatrix} + \begin{bmatrix} y_1, t-1 \\ y_2, t-1 \end{bmatrix} + ... + \begin{bmatrix} \phi_{11}\phi_{12}.. \\ \phi_{21}\phi_{22}.. \end{bmatrix} + \begin{bmatrix} y_1, t-p \\ y_2, t-p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

Where y is the endogenous variable, phi is the coefficient of lags of y, p is the number of lags and epsilon is the error term.

Just as with the AR models, forecasts using both, monthly and daily datasets were calculated. Lagged variables of the CPI had been used alongside the shipping data and some of the macroeconomic and market variables as previously used in the literature (D'Agostino's et al, 2009 and Faust and Wright, 2011) [11] [17]. For the VAR model, CPI differences, Non-Farm Pay Roll (unemployment), ISMM (PMI), average ship waiting time, and daily number of ships in the top 40 ports were used as variables. Data was checked for stationarity using the AD Fuller test and first differences were taken when appropriate. The optimum number of lags to be used, was determined by calculating the AIC score across numerous lags and selecting the one with the lowest AIC score. Subsequently, VAR model was fitted on the same training data period as with the AR model. The last 12 months were then used as a test set to calculate the MSE, which was compared to other models. All the ADF, AIC, and VAR results can be found in Appendix F.

### 4.3 Machine Learning Models

Random Forest as described by Breiman (2001) was used as a ML method in this study. This is a supervised machine learning technique that initially trains on known outputs and is then tested with previously unseen testing data. Random Forest algorithm works by combining many decision trees in an ensemble, in an uncorrelated fashion to obtain an overall result. This methods makes it be more accurate than that of any individual decision tree [5].

The decision trees themselves work by splitting data at specific nodes as demonstrated in Figure 8. The data is "sorted" once certain "thresholds" are breached. These "thresholds" are determined during the training step of the model.

In the Random Forest, numerous decision trees are employed using the ensemble method that involves boosting and bootstrap aggregation, otherwise known as bagging (Breiman, 1996). In this method, different trees are trained on randomly selected data with each tree producing its own output. Once the outputs are averaged out, an overall global output is obtained. Random Forest models take it a step further by creating a set of uncorrelated decision trees. This is a crucial step that allows Random Forest to reduce variance thereby reducing the risk of overfitting and which is why this method has become so extensively used in research [4].

Random Forest method used in this study is obtained from the "Sklearn" Python library and is automatically pre-set with default hyperparameters. The main hyperparameters include: the number of trees, maximum depth of the tree, minimum number of samples required to split, and maximum number of features to consider when looking for a best split. These hyperparameters are tuned using the GridSearch method, which finds the combination of hyperparameters that produces the smallest MSE, however this was not within the scope of the study, but should be looked into in the future.

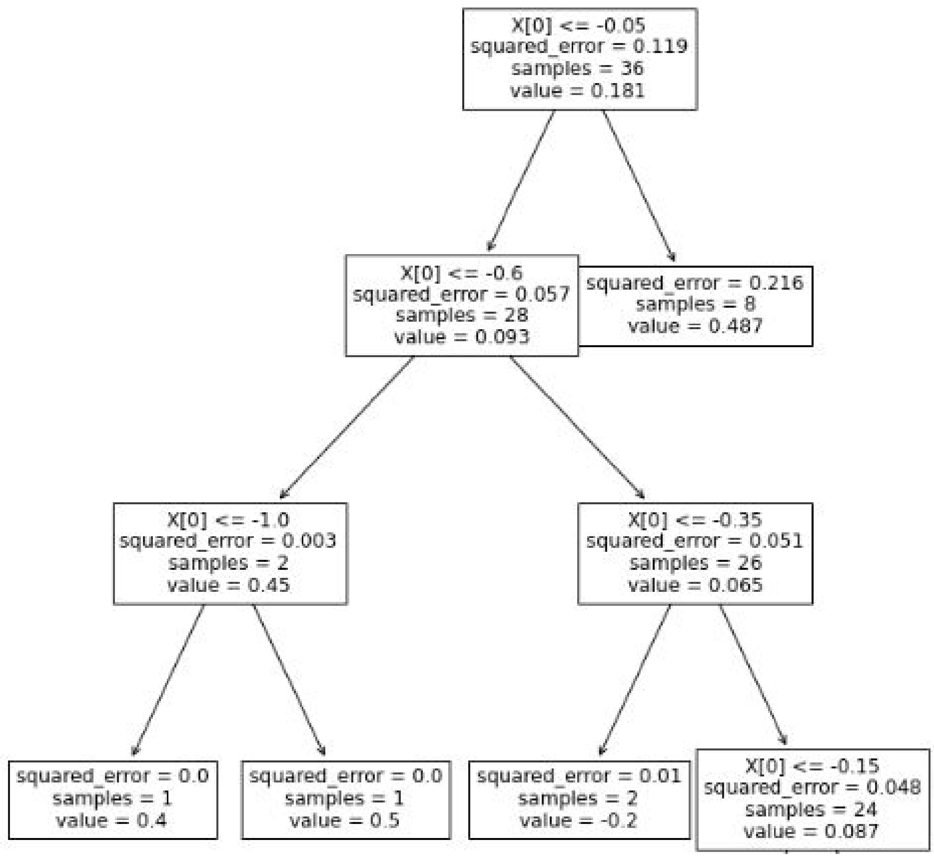


Figure 8: Decision Tree of the AR ML model. X[0] represents the value of the lagged CPI variable.

For this study, the first Random Forest model was trained only using lagged CPI differences as a "benchmark" ML model to compare to the AR and final model results. Then, other variables were added, starting with the shipping data, followed by market and economic data. As overall, there were 28 variables in the dataset (14 derived from shipping and 14 from market data), an algorithm was created to calculate an MSE of different variable combinations. The algorithm works by initially running the models using one variable, then a combination of two, and so on. Due to computational limitations, datasets were split into "Shipping" and "Shipping and Market", where 16,141 model combinations were ran for each set. The combination with the lowest MSE was chosen as the final model. It was also decided that the CPI lagged variable would be present in all models as from literature this appeared to be the best performing feature. As with the AR benchmarks, both monthly and daily models were optimised. Performance was monitored by comparing the MSE scores.

The appropriate lags of the shipping variables had to also be determined. It seemed logical that the effects of shipping delays would not instantaneously affect the economy and the inflation figures. Preliminary results, using only the shipping data whilst changing lag lengths, showed that lag of 4 months produced the best results (Table 6). For the market variables, shorter lag of one month was chosen as it is believed that markets are normally able to reflect all the available information much more rapidly.

| Lags | MSE |
|---|---|
| 2 | 0.2109 |
| 3 | 0.2015 |
| 4 | 0.125*** |
| 5 | 0.1526 |
| 6 | 0.1436 |

Table 6: Preliminary test results to determine the best Lag. *** lowest MSE (optimum lag).

### 4.4 Diebold Mariano Significance Test

Once the most optimum models were established, the statistical significance of differences between the benchmark and the final model was calculated. The lagged uni-variate (lagged CPI) ML model was used as a comparison benchmark. This was posed against the four best models from the four different datasets used in the study. To calculate the significance, the Diebold-Mariano test was used. This test appeared to be best suited for comparing MSE scores from different models with non-Gaussian error terms (Diebold and Mariano, 2002) [15]. The python code for this test was found at: https://github.com/johntwk/Diebold-Mariano-Test/blob/master/dm_test.py.

### 4.5 Feature Importance and Explainability

Finally, the best performing models were analysed using shapley values in order to explore and attempt to explain the model. The "SHAP" Python package is able to show how each feature contributes to the final output on both, global and local scales https://shap.readthedocs.io/en/latest/. Based on the cooperative game theory approach, shapley values are calculated for each feature which indicate how they contributed to the final output.

## 5 Results

### 5.1 Counterfactual Inflation Impact Estimation

The counterfactual study performed in this research demonstrated that the increase in the number of ships at the top 40 US ports had a significant impact on the inflation. Both exogenous variables were statistically significant in the OLS regression. Results indicate that every additional vessel to the average number of ship at the top 40 ports contributes 0.0042% to the CPI figure (full OLS results are presented in Appendix G). The model estimates that if the number of ships at the US ports remained

at the pre-pandemic average, the CPI figure would have been approximately 6.34% instead of 8.6% by April of 2022 (Figure 9). At the start of the period, the "no bottleneck" inflation actually appears to be higher as the number of ships goes down as a result of border closures and restrictions on movement. However, the picture changes by the end of the year where the number of ships at the ports rises significantly compared to the pre-pandemic average.

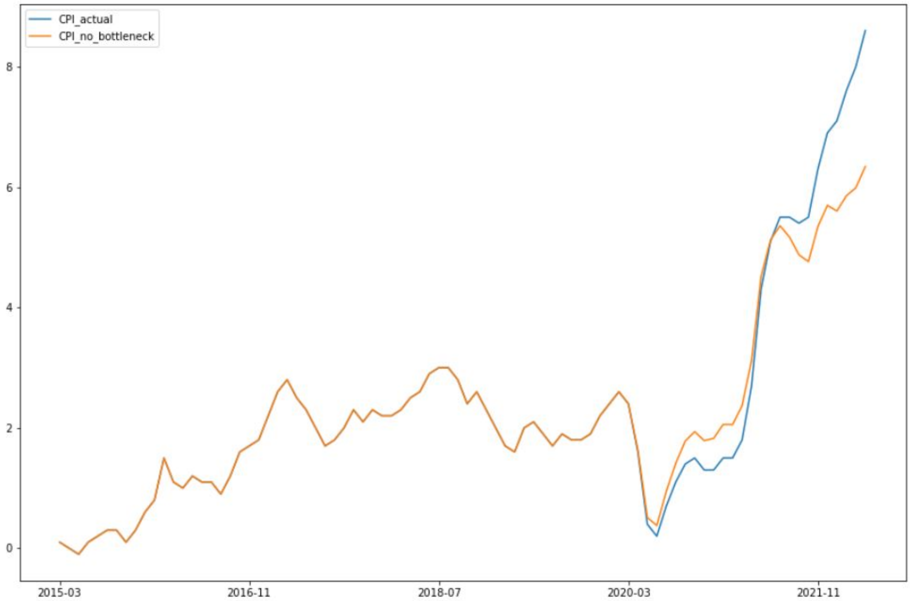


Figure 9: Counterfactual Inflation impact Estimation if number of ships at top 40 ports remained at the prior to March 2020 levels.

## 5.2 Benchmarks

### 5.2.1 Autoregressive Benchmark Model

The AR benchmarks had a very respectable low MSE of 0.131 for the monthly data and 0.116 for daily data over a 10 months testing period. As can be seen in Figure 10, AR models perform extremely well on short period horizon, however start to perform poorly on long term as the values revert to the average. The OLS results show that for monthly data, first lag was significant, and for daily the first four lags were significant, results presented in Table 7 and Appendix F.

| | **coef** | **std err** | **z** | **P> \|z\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **intercept** | 0.0418 | 0.036 | 1.159 | 0.246 | -0.029 | 0.113 |
| **CPI_diff.L1** | 0.6483 | 0.126 | 5.148 | 0.000 | 0.401 | 0.895 |
| **CPI_diff.L2** | -0.2440 | 0.133 | -1.830 | 0.067 | -0.505 | 0.017 |

Table 7: AR OLS coefficients and p values.

### 5.2.2 Vector Autoregressive Benchmark Model

For the monthly data, maximum order was determined to be two using the AIC criteria. The VAR CPI model showed significance for four variables: first lag of the CPI difference, and the first and second lags of daily number of ships at the top busiest ports (Table 8). For the daily dataset, the optimum number of lags was also determined to be two, with significance observed in the first and second lags of CPI and Non-Farm Payroll (Table 9 and Appendix F). The overall 12 months forecast had an MSE of 0.199 for the monthly and 0.117 for daily datasets and is presented in Figure 11.

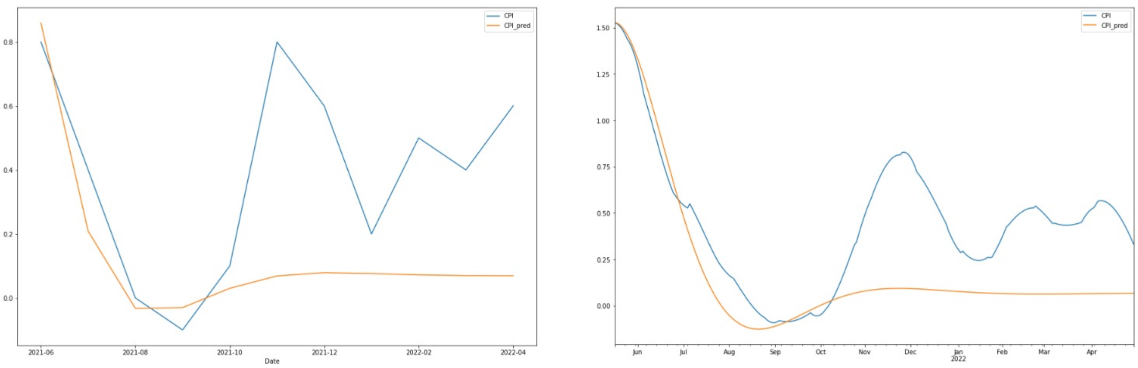


Figure 10: AR Benchmark forecast v actual - monthly (left) and daily (right).

| | coef | std err | T-stat | prob |
|---|---|---|---|---|
| **L1.CPI** | 0.514 | 0.158 | 3.237 | 0.001 |
| **L1.Mean Top40** | 0.010 | 0.005 | 1.752 | 0.08 |
| **L2.Mean Top40** | -0.010 | 0.005 | -1.949 | 0.051 |

Table 8: Monthly VAR coefficients and p values.

| | coef | std err | T-stat | prob |
|---|---|---|---|---|
| **L1.CPI** | 1.961 | 0.007 | 299.243 | 0.000 |
| **L1.NFPTOT** | 0.000003 | 0.000001 | 2.335 | 0.020 |
| **L2.CPI** | -0.964 | 0.007 | -146.953 | 0.000 |
| **L2.NFPTOT** | -0.000003 | 0.000001 | -2.368 | 0.018 |

Table 9: Daily VAR coefficients and p values.

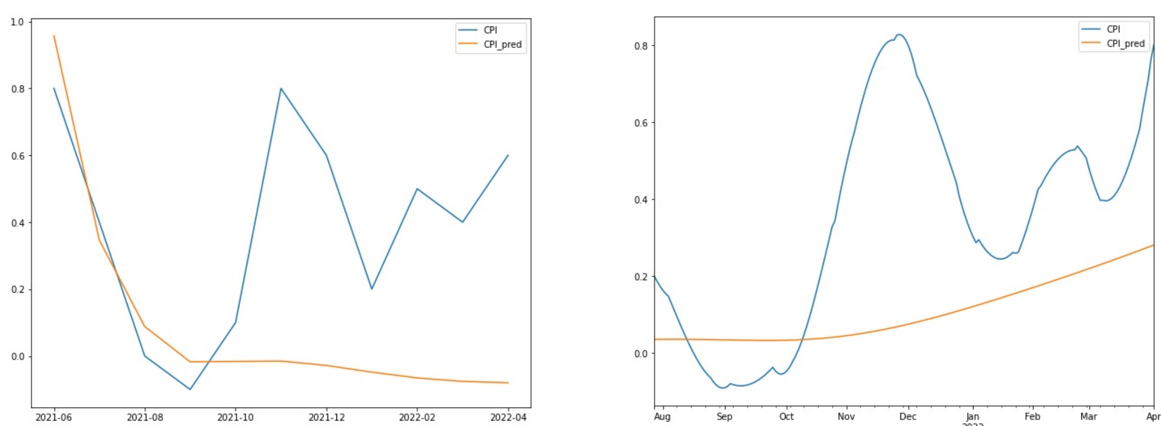


Figure 11: VAR Benchmark forecast v actual - monthly (left) and daily (right).

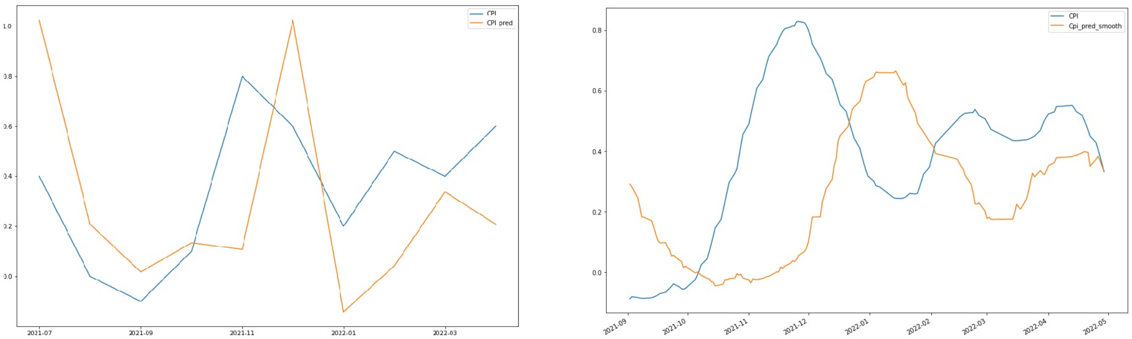


Figure 12: ML Benchmark forecast v actual - monthly (left) and daily (right).

### 5.2.3 Machine Learning Benchmark Model

The ML benchmarks using random forest and only lagged CPI figures as exogenous variables showed similar results for monthly model - MSE of 0.132 but slightly worse performance for the daily - MSE of 0.149, as seen in Table 10 and in Figure 12.

| | **Monthly** | **Daily** |
|---|---|---|
| AR models | 0.131 | 0.116 |
| VAR Models | 0.199 | 0.117 |
| ML models | 0.132 | 0.149 |

Table 10: AR and ML benchmark MSE for daily and monthly models.

## 5.3 Random Forest Machine Learning Models

### 5.3.1 Monthly Data - Shipping Data Only

Using the 14 shipping variables calculated from the AIS data, 16,384 combinations of these variables were calculated and analysed using the Random Forest models. MSE ranged from 0.072 to 0.247, summary statistics are presented in Table 15 and best and worst combinations with the lowest and highest MSEs are in Table 17. Table with best 10 and worst 10 combinations are also presented in the Appendix E. Graphs of the best and worst performing models are shown in Figure 13.

| | **Features** | **MSE** |
|---|---|---|
| **BEST** | CPI LAG, Miami, Port Everglades, Bayonne | 0.072 |
| **WORST** | CPI LAG, Mean Number, Mean Top40, Norsworthy, Port Everglades, Port Time, Bayonne, Galveston, Alameda, Savannah | 0.247 |

Table 11: Best and Worst feature combinations using Monthly Shipping Data.

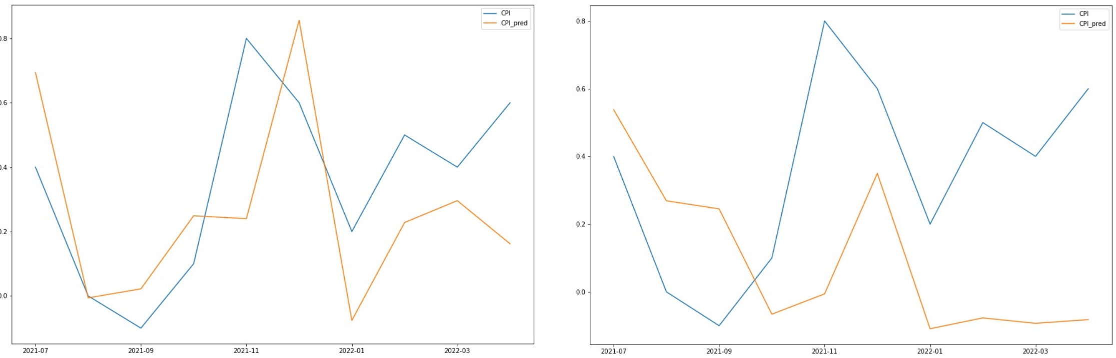


Figure 13: Best (MSE: 0.072) and Worst (MSE: 0.247) Performing Monthly Shipping models.

#### 5.3.2 Monthly Data - Shipping and Market Data

The features that produced the best results in the previous section were kept and new market variables were added. With new market indicators MSE had improved, and ranged from 0.048 to 0.495, summary statistics are presented in Table 15 The best and worst combinations with lowest and highest MSEs are in Table 22. Tables with the best 10 and worst 10 combinations are also presented in the Appendix E. Graphs of the best and worst performing models are shown in Figure 14.

| | Features | MSE |
|---|---|---|
| **BEST** | CPI LAG, Bayonne, TIPS10, Alameda | 0.048 |
| **WORST** | CPI LAG, Bayonne, Galveston, GOLDCOPPER, CRB | 0.495 |

Table 12: Best and Worst feature combinations using Monthly Shipping and Market Data.

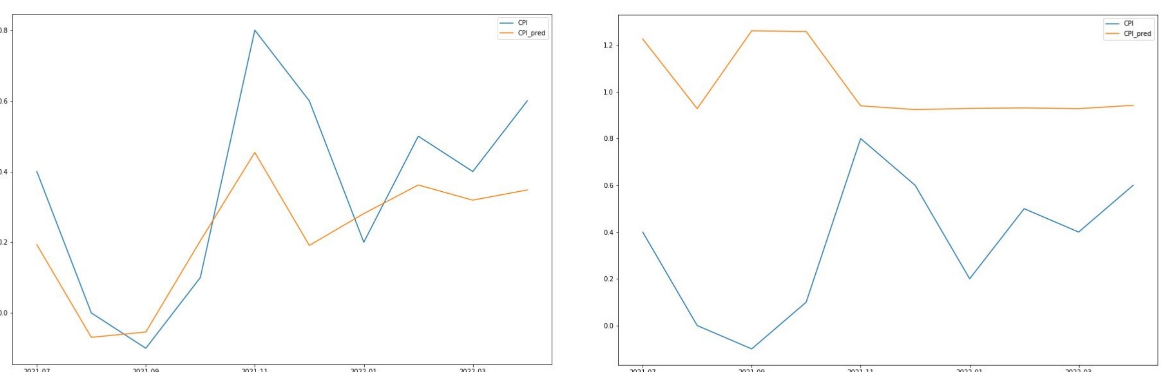


Figure 14: Best (MSE: 0.048) and Worst (MSE: 0.495) Performing Monthly Shipping and Market models .

#### 5.3.3 Daily Data - Shipping Data Only

Due to limitations with computing powers, only 8,191 combinations of the shipping daily variables were ran. For these models, MSE ranged from 0.056 to 0.341, summary statistics are presented in Table 15. The best and worst combinations with lowest and highest MSEs are in Table 21. Tables with the best 10 and worst 10 combinations are also presented in the Appendix E. Graphs of the best and worst performing models are shown in Figure 15.

| | **Features** | **MSE** |
|---|---|---|
| **BEST** | Mean Top40, Bayonne(T), Alameda(T), Port Everglades(T), CPI Lag | 0.062 |
| **WORST** | Bayonne(T), Alameda(T), Savannah(T), CPI Lag | 0.341 |

Table 13: Best and Worst feature combinations using Daily Shipping Data.

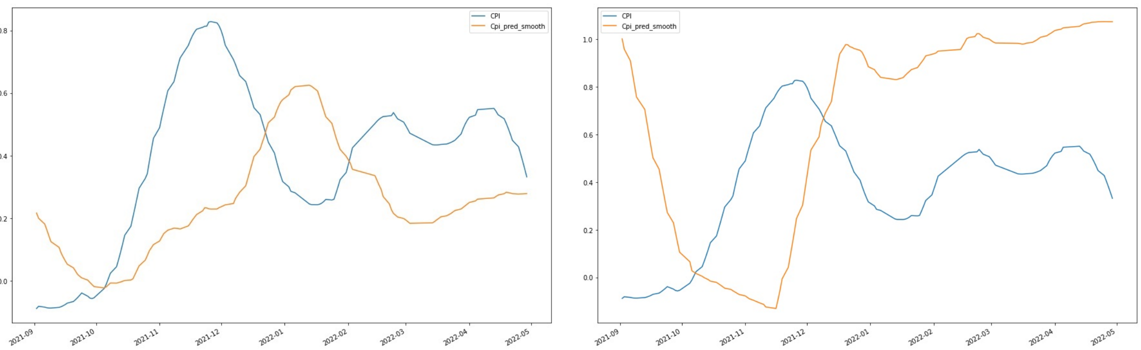


Figure 15: Best (MSE: 0.062) and Worst (MSE: 0.341) Performing Daily Shipping models.

#### 5.3.4 Daily Data - Shipping and Market Data

Just as with the monthly data, features that produced the best results using only daily "Shipping Data", were kept and additional market variables were added. However, even with new market indicators, the best performing model did not change. The lowest MSE stayed at 0.062 using the same combination of variables as used in the "Shipping Only model", summary statistics are presented in Table 15. The best and worst combinations with lowest and highest MSEs are in Table 23. Tables with the best 10 and worst 10 combinations are also presented in the Appendix E. Graphs of the best and worst performing models are shown in Figure 16.

| | **Features** | **MSE** |
|---|---|---|
| **BEST** | Bayonne(T), Mean Top40, Alameda(T), Port Everglades(T), CPI Lag | 0.062 |
| **WORST** | Mean Top40, Gold/Copper, InflationSwap5Y5Y, Port Everglades(T), CPI Lag | 0.427 |

Table 14: Best and Worst feature combinations using Daily Shipping and Market Data.

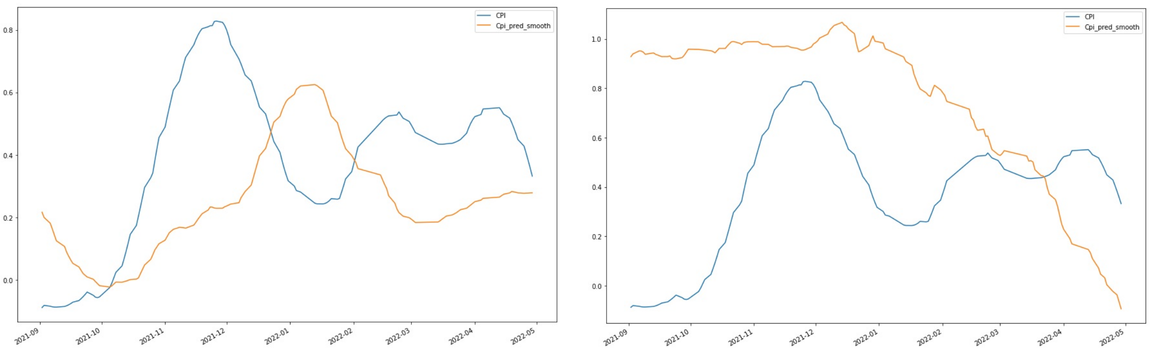


Figure 16: Best (MSE: 0.062) and Worst (MSE: 0.427) Performing Daily Shipping and Market models.

| | MSE (Monthly Ship) | MSE (Monthly Ship/Market) | MSE (Daily Ship) | MSE (Daily Ship/Market) |
|---|---|---|---|---|
| **mean** | 0.151 | 0.237 | 0.166 | 0.207 |
| **std** | 0.028 | 0.090 | 0.040 | 0.069 |
| **min** | 0.072 | 0.048 | 0.062 | 0.062 |
| **25%** | 0.131 | 0.158 | 0.135 | 0.144 |
| **50%** | 0.148 | 0.249 | 0.163 | 0.205 |
| **75%** | 0.169 | 0.301 | 0.196 | 0.258 |
| **max** | 0.247 | 0.495 | 0.341 | 0.427 |

Table 15: Summary Statistics of MSE scores for all the variables combinations for the four datasets.

#### 5.3.5 Diebold-Mariano Significance test

The results of the Diebold-Mariano test are presented in Table 16. This test compared the ML benchmark MSE to the final model MSE. For the monthly models, the significance was established to be at just over 94% for both, Shipping Only and Shipping and Market datasets. For the Daily models, the significance was over 99% for both datasets as the model remained unchanged.

| ML Model | MSE | p-value | Bench % diff |
|---|---|---|---|
| **AR (Monthly Benchmark)** | 0.132 | NA | NA |
| **AR (Daily Benchmark)** | 0.149 | NA | NA |
| **Monthly Ship Only** | 0.072 | 0.0599 | 83.3% |
| **Monthly Ship + Market** | 0.048 | 0.0596 | 175% |
| **Daily Ship Only** | 0.062 | 0.000 | 140% |
| **Daily Ship + Market** | 0.062 | 0.000 | 140% |

Table 16: Statistical Significance of ML models compared to Benchmarks using Diebold-Mariano Test.

### 5.4 Machine Learning Model Explainability

The feature importance graphs of all models, as calculated and visualised by the "SHAP" package, are presented in Figures 17, 18, and 19. Unsurprisingly, the lagged CPI variable is the most important feature in all the three models, this is most evident in the "Daily model". For the "Monthly models", inclusion of the Treasury Inflation-Protected Security 10 year yield (TIPS10) was also significantly important, as was the number of ships at the ports of Alameda and Bayonne. Without TIPS10 variable, the number of ships at Baytown, Port Everglades and Miami were the most important. In the "Daily model", the waiting times at Alameda, Bayonne, and Port Everglades were considered almost equally important.

The calculated Shapley values suggest that for the "Monthly models", the high lagged CPI numbers contribute to an increase in CPI for the next month and low lagged CPI numbers decrease the CPI. In the "Shipping Only" model, for the port of Baytown, the high number of ships waiting to be unloaded contribute positively to CPI and negatively when the number is low. Similarly this was observed in the port of Miami, albeit to a lesser extent. Surprisingly, the Shapley values for the port of Everglades were not consistent as high values appeared to sometimes to be linked with increases and decreases in CPI, with occasions where lower number of ships appears to influence increases in CPI.

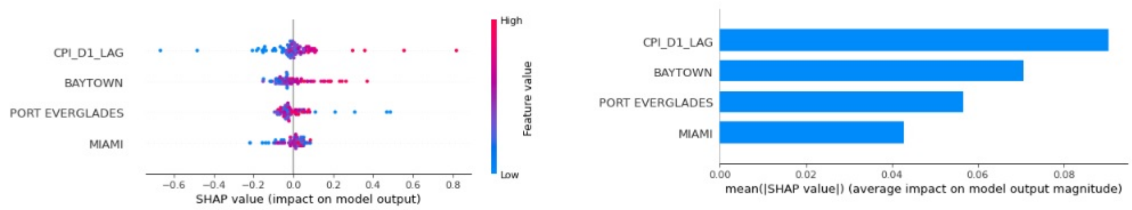


Figure 17: Shapley Value Plot and Feature Importance of the Best Monthly Shipping Data Only Model.

For the "Shipping and Market Monthly model", the Shapley values provide a much clearer picture. As previously mentioned, higher lagged CPI values indicate an increase in CPI, however the newly included TIPS10 yields suggest that higher rates indicate decrease in CPI amd visa versa. The higher numbers of ships at ports of Alameda and Bayonne also suggest to be positively affecting CPI changes.

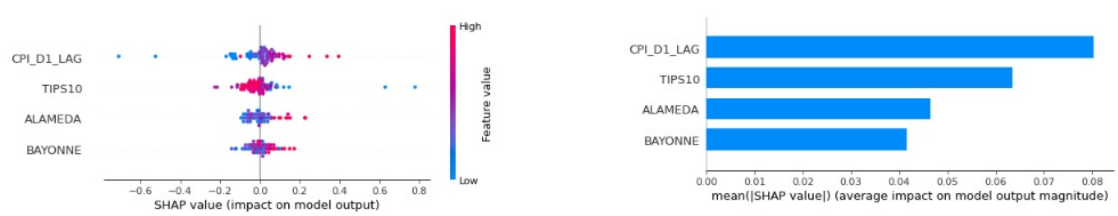


Figure 18: Shapley Value Plot and Feature Importance of the Best Monthly Shipping and Market Data Model.

In the "Daily model", the Shapley value diagram becomes more detailed due to the increased amount of observations. Just as with the "Monthly models", the diagram shows that higher lagged CPI values indicate a higher chance for the next CPI to be higher. The time spent at the Alameda port and the mean number of ships at the Top 40 ports appear to be inconsistent with both, high and low values of these variables having instances where they contribute positively and negatively towards the predicted

CPI change. On the other hand, the time spent by the ships at the ports of Bayonne and Everglades seem to suggest that a delay will positively affect the next CPI number and visa versa.

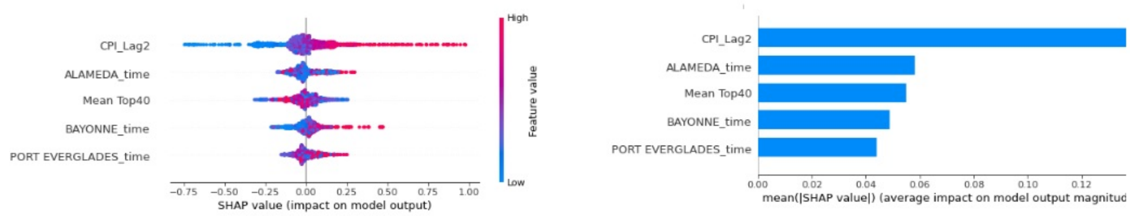


Figure 19: Shapley Value Plot and Feature Importance of the Best Daily Shipping Only Model (The Same model was the best performer for the Shipping and Market Data Only).

## 6 Discussion

The results of this study show that the number of ships at the ports of US do have an impact on the inflation and Machine Learning models that use AIS shipping data categorically outperform the Benchmarks and could potentially be even further improved using additional market data. The counterfactual study estimated that inflation in the US would have been over 2.2% lower by April 20222, had the number of ships at the ports remained at the pre-pandemic levels. For the forecasting models, all showed statistical significance when tested with Diebold-Mariano against the ML benchmark with outperformance of between 83% and 175% according to their MSE scores.

As previously mentioned in the introduction to this paper, the improvements can be rationalised by looking at the supply side of the US economics. As US/UK Economists have previously suggested, current rise in inflation had started when demand remained at the same or increased levels, whilst the supply decreased as a result of the pandemic [14] and [22],[27] . Since 80% of all goods arrive to the US by sea, any disruptions to these maritime supply chains will likely to have significant repercussions on the rest of the economy. Specifically, as goods struggle to reach the final consumer, it pushes prices up as demand outstrips the supply thereby increasing inflation. According to the AIS data analysed, the average number of ships waiting at the ports increased from 200 at the beginning of 2020 to 300 by the end of 2021, with average waiting times tripling in ports like Alameda in San Francisco from 40 to 120 hours. These delays could explain why the shipping data had improved the models that were used in this study. However, it is necessary to point out that according to the feature importance graphs, the lagged CPI variable remained the most important feature of the models.

When we look deeper into the best performing models, we are able to distinguish which variables appear to be the most significant. Purely from absolute point of view, we see that the number of ships at port of Baytown (Texas) appeared in 8 out of 10 best performing monthly models, followed by ports of Maimi (Florida) and Alameda (California) with 5 out of 10 appearances. These ports represented 1.37%, 0.04%, and 0.12% of all US trade in 2013 respectively. As Baytown is the "busiest" port in the US in terms of trade, it is safe to say that any disturbances in this port could indeed cause a shock that is significant to the economy. This is further supported by the SHAP explainability plot that clearly shows high number of ship at the port contribute to higher CPI and visa versa. On the other hand, port of Everglades also appeared 8 out of 10 times in the best models, but 10 out of 10 times in the worst monthly models, pointing towards insignificance of this port data as a variable or inconsistency of the models. The inconsistency with this variable is also seen in the SHAP plot that seemed to indicate that high values contribute to both, increases and decreases in CPI.

Additional market data did improve the MSE scores with the introduction of the Treasury Inflation-Protected Securities (TIPS10) in all of the top 10 monthly models. The price of Oil also appeared in the 3 out of 10 models. According to the Shapley values, lower TIPS rates predict increase in inflation. To some extent this is reasonable, as since 2020 the yields on the TIPS have been negative suggesting that markets have already "priced in" the expected higher inflation into the bond price that in itself implies higher future inflation that is potentially identified by the model.

Similar conclusions can be drawn from the "Daily model". Lagged CPI variable was the most important feature, with higher values predicting increases in CPI. Increase in the waiting times at the port of Baytown were sh own to also increase the likelihood of CPI increase. The other variables are not as consistent according to their Shapley values. The idea behind including a daily model was to increase the amount of observations on which the ML model can be trained, thereby increasing its accuracy, however from the results obtained it is not convincing that the model had improved significantly. Firstly, the MSE scores were lower than that of the best "Monthly model" and secondly, the explainability was not able to provide much more clarity on the decision making process.

In addition to only gaining limited improvements using daily data, it is necessary to recognise other limitations of this study. Even though this study shows that shipping data is a useful variable to forecast inflation, it is very important to look at the test and training periods of the model and the specific economic environment during this time. The COVID pandemic had created a unique and unprecedented global situation that completely disrupted the supply chains causing, at least initially, this demand-pull inflation. Without further data that captures similar periods of rapid inflation growth, as was seen in the 1980s, it would be difficult to assess if these model would still hold up. There are reasons to be doubtful that they would, especially as the current supply chains have never been as globally intertwined as ever before. As of 2019, the KOF Globalisation Index had reached an all time high of 62 as compared to 40 in the 1980s [20], suggesting that any shocks to the supply chains anywhere in the world would have a much greater ripple effect now than in previous years.

# 7 Conclusion

This study has contributed to the existing literature by demonstrating that using alternative data, such as the Automatic Identification System used in vessel tracking, in combination with Machine Learning techniques, considerably improves the performance of inflation forecasting models. Additionally, this study adds to literature by utilising explainable AI (SHAP package) to be able to interpret the decision making processes of the models.

The performance of these models however, needs to be viewed in context. Even though the models had outperformed the benchmarks by a significant percentage, it is necessary to note that these results were obtained over a very unprecedented economic and social period. The pandemic had put an additional strain on the supply chains which, according to the consensus of the majority of the economists, had caused a negative supply shock whilst the demand remained strong. This theory was further backed by the results of the counterfactual study, that demonstrated that supply chain bottlenecks had contributed approximately 2.2% to the inflation over the two years since March 2020. Previous periods in history of high inflation were not characterised by such supply shocks, therefore it is questionable how well the study's model would have performed during previous periods. It is hard to imagine that during the period of high inflation in 1980s, the AIS data would have been useful for forecasting, as the inflation during that period was mainly driven by high commodity prices. Unfortunately, the AIS data was only available for the previous seven years and therefore it was not possible to confirm the validity of the model for other time frames. In this study, data availability was the main limiting factor and therefore ideally would need to be repeated with a larger dataset.

According to literature, larger datasets are also more preferred when using the Machine Learning techniques. In this study, the number of observations was increased by upsampling to daily frequency. However, any significant changes in the models were not observed when monthly and daily models were compared.

Finally, the Machine Learning explainability aspect of the study showed promising results. Each feature of the model was analysed and it was possible to see exactly how much of an impact they had to the final output. The SHAP visualisation showed that the busiest port of the USA was indeed contributing to the inflation figure when the delays increased. On the other hand, other ports appeared to not to be significant and using this information the model can be further optimised. Perhaps the most unsurprising conclusion of the explainability part of the study was that the lagged CPI values remained the most important variable confirming the findings in the literature of the last 40 years.

# References


[1] Andrew Atkeson, Lee E Ohanian, et al. Are phillips curves useful for forecasting inflation? *Federal Reserve bank of Minneapolis quarterly review*, 25(1):2–11, 2001.

[2] Ben S Bernanke and Michael Woodford. Inflation forecasts and monetary policy, 1997.

[3] Dillon Bowen and Lyle Ungar. Generalized shap: Generating multiple types of explanations in machine learning. *arXiv preprint arXiv:2006.07155*, 2020.

[4] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] John Y Campbell and Robert J Shiller. A scorecard for indexed government debt. *NBER macroeconomics annual*, 11:155–197, 1996.

[7] John Y Campbell, Robert J Shiller, and Luis M Viceira. Understanding inflation-indexed bond markets. Technical report, National Bureau of Economic Research, 2009.

[8] Mr Yan Carriere-Swallow, Mr Pragyan Deb, Davide Furceri, Daniel Jimenez, and Mr Jonathan David Ostry. *Shipping Costs and Inflation.* Number 17259. International Monetary Fund, 2022.

[9] Alberto Cavallo. Scraped data and sticky prices. *Review of Economics and Statistics*, 100(1):105–119, 2018.

[10] Todd E Clark. Real-time density forecasts from bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics*, 29(3):327–341, 2011.

[11] Antonello D'Agostino, Luca Gambetti, and Domenico Giannone. Macroeconomic forecasting and structural change. *Journal of applied econometrics*, 28(1):82–101, 2013.

[12] Marcos Lopez De Prado. *Advances in financial machine learning*. John Wiley & Sons, 2018.

[13] Olivier Dessaint, Thierry Foucault, and Laurent Frésard. Does alternative data improve financial forecasting? the horizon effect. 2021.

[14] Julian Di Giovanni, ebnem Kalemli-Özcan, Alvaro Silva, and Muhammed A Yildirim. Global supply chain pressures, international trade, and inflation. Technical report, National Bureau of Economic Research, 2022.

[15] Francis X Diebold and Robert S Mariano. Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144, 2002.

[16] Eugene F Fama. Efficient market hypothesis. *Diss. PhD Thesis, Ph. D. dissertation*, 1960.

[17] Jon Faust and Jonathan H Wright. Forecasting inflation. In *Handbook of economic forecasting*, volume 2, pages 2–56. Elsevier, 2013.

[18] Paolo Fornaro and Henri Luomaranta. Nowcasting finnish real economic activity: a machine learning approach. *Empirical Economics*, 58(1):55–71, 2020.

[19] Alan Greenspan. Statements to the congress. *Fed. Res. Bull.*, 78:329, 1992.

[20] Savina Gygli, Florian Haelg, Niklas Potrafke, and Jan-Egbert Sturm. The kof globalisation index–revisited. *The Review of International Organizations*, 14(3):543–574, 2019.

[21] R. W. Hafer and Scott E. Hein. Forecasting inflation using interest-rate and time-series models: Some international evidence. *The Journal of Business*, 63(1):1, 1990.

[22] Jesse LaBelle and Ana Maria Santacreu. Global supply chain disruptions and inflation during the covid-19 pandemic. *Federal Reserve Bank of St. Louis Review*, 2022.

[23] Burton G Malkiel. Efficient market hypothesis. In *Finance*, pages 127–134. Springer, 1989.

[24] Marcelo C Medeiros, Gabriel FR Vasconcelos, Álvaro Veiga, and Eduardo Zilberman. Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119, 2021.

[25] Christoph Molnar, Timo Freiesleben, Gunnar König, Giuseppe Casalicchio, Marvin N Wright, and Bernd Bischl. Relating the partial dependence plot and permutation feature importance to the data generating process. *arXiv preprint arXiv:2109.01433*, 2021.

[26] Matthieu Picault, Julien Pinter, and Thomas Renault. Media sentiment on monetary policy: determinants and relevance for inflation expectations. *Journal of International Money and Finance*, 124:102626, 2022.

[27] Ricardo Reis. The burst of high inflation in 2021–22: How and why did we get here? 2022.

[28] Ronald W Schafer. What is a savitzky-golay filter?[lecture notes]. *IEEE Signal processing magazine*, 28(4):111–117, 2011.

[29] James H Stock and Mark W Watson. Why has us inflation become harder to forecast? *Journal of Money, Credit and banking*, 39:3–33, 2007.

[30] James H Stock and Mark W Watson. Modeling inflation after the crisis. Technical report, National Bureau of Economic Research, 2010.

# Appendices

## A Raw AIS and WPI data Examples

| MMSI | BaseDateTime | LAT | LON | SOG | COG | Heading | VesselName | IMO | CallSign | VesselType | Status | Length | Width | Draft | Cargo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 367702220 | 2022-03-31 00:00:01 | 29.78763 | -95.08070 | 0.1 | 226.5 | 340.0 | JOE B WARD | NaN | WDI4808 | 31.0 | 12.0 | 21.0 | 8.0 | NaN | 57.0 |
| 671226100 | 2022-03-31 00:00:01 | 25.77626 | -80.20320 | 3.2 | 143.7 | 511.0 | RELIANCE II | IMO9221322 | 5VHS7 | 79.0 | 0.0 | 52.0 | 12.0 | 2.5 | 70.0 |
| 367767250 | 2022-03-31 00:00:01 | 29.31623 | -94.78829 | 4.5 | 228.1 | 511.0 | GLEN K | NaN | WDJ3358 | 52.0 | 0.0 | 0.0 | 0.0 | 0.0 | 52.0 |
| 338327436 | 2022-03-31 00:00:03 | 47.29634 | -122.42233 | 0.0 | 360.0 | 511.0 | COOL KAT | IMO0000000 | NaN | 36.0 | NaN | 15.0 | 3.0 | NaN | NaN |
| 367452810 | 2022-03-31 00:00:06 | 29.32824 | -94.77391 | 2.6 | 319.2 | 511.0 | JOHN W JOHNSON | IMO9602344 | WDF4516 | 60.0 | 0.0 | 80.0 | 19.0 | 3.0 | 60.0 |

Figure 20: Example of raw AIS (Automatic Identification System) data with all the available information.

| REGION_NO | PORT_NAME | COUNTRY | LATITUDE | LONGITUDE | LAT_DEG | LAT_MIN | LAT_HEMI | LONG_DEG | ... | ELECREPAIR | PROVISIONS | WATER | FUEL_OIL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 61070 | SHAKOTAN | RU | 43.866667 | 146.833333 | 43 | 52 | N | 146 | ... | NaN | NaN | NaN | NaN |
| 61100 | MOMBETSU KO | JP | 44.350000 | 143.350000 | 44 | 21 | N | 143 | ... | NaN | NaN | Y | NaN |
| 61100 | ABASHIRI KO | JP | 44.016667 | 144.283333 | 44 | 1 | N | 144 | ... | NaN | NaN | Y | NaN |
| 61100 | NEMURO KO | JP | 43.333333 | 145.583333 | 43 | 20 | N | 145 | ... | NaN | Y | Y | Y |
| 61100 | HANASAKI KO | JP | 43.283333 | 145.583333 | 43 | 17 | N | 145 | ... | NaN | NaN | Y | Y |

Figure 21: Example of raw WPI (World Port Index) data with all the available information.

| | | | | |
|---|---|---|---|---|
| | Passenger | 60 | 60 | Passenger, all ships of this type |
| | Passenger | 61 | 61 | Passenger, hazardous category A |
| | Passenger | 62 | 62 | Passenger, hazardous category B |
| | Passenger | 63 | 63 | Passenger, hazardous category C |
| | Passenger | 64 | 64 | Passenger, hazardous category D |
| | Passenger | 65 | 65 | Passenger, reserved for future use |
| | Passenger | 66 | 66 | Passenger, reserved for future use |
| | Passenger | 67 | 67 | Passenger, reserved for future use |
| | Passenger | 68 | 68 | Passenger, reserved for future use |
| | Passenger | 69 | 69 | Passenger, no additional information |
| | Cargo | 70 | 70 | Cargo, all ships of this type |
| | Cargo | 71 | 71 | Cargo, hazardous category A |
| | Cargo | 72 | 72 | Cargo, hazardous category B |
| | Cargo | 73 | 73 | Cargo, hazardous category C |
| | Cargo | 74 | 74 | Cargo, hazardous category D |
| | Cargo | 75 | 75 | Cargo, reserved for future use |
| | Cargo | 76 | 76 | Cargo, reserved for future use |
| | Cargo | 77 | 77 | Cargo, reserved for future use |
| | Cargo | 78 | 78 | Cargo, reserved for future use |
| | Cargo | 79 | 79 | Cargo, no additional information |
| | Tanker | 80 | 80 | Tanker, all ships of this type |
| | Tanker | 81 | 81 | Tanker, hazardous category A |
| | Tanker | 82 | 82 | Tanker, hazardous category B |
| | Tanker | 83 | 83 | Tanker, hazardous category C |

Figure 22: Vessel type codes example, used to clean AIS data.

| Status Code | Description |
|---|---|
| 0 | Under way using its engine |
| 1 | Anchored |
| 2 | Not under command |
| 3 | Has restricted maneuverability |
| 4 | Ship draught is limiting its movement |
| 5 | Moored (tied to another object to limit free movement) |
| 6 | Aground |
| 7 | Engaged in fishing |
| 8 | Under way sailing |
| 9 | (Number reserved for modifying reported status of ships carrying dangerous goods/harmful substances/marine pollutants) |
| 10 | (Number reserved for modifying reported status of ships carrying dangerous goods/harmful substances/marine pollutants) |
| 11 | Power-driven vessel towing astern |
| 12 | Power-driven vessel pushing ahead/towing alongside |
| 13 | (Reserved for future use) |
| 14 | Any of the following are active |
| 15 | Undefined (default) |

Figure 23: Vessel status codes example, used to clean AIS data.

## B Summary Statistics (Monthly Data)

| | NORSWORTHY | PORT EVERGLADES | BAYONNE | GRETNA | GALVESTON |
|---|---|---|---|---|---|
| count | 82.000000 | 82.000000 | 82.000000 | 82.000000 | 82.000000 |
| mean | 14.370567 | 9.057525 | 8.573390 | 6.489270 | 3.382475 |
| std | 2.199803 | 1.452229 | 1.276950 | 1.379807 | 0.879840 |
| min | 9.483333 | 6.020161 | 5.897335 | 2.728210 | 1.752676 |
| 25% | 13.037258 | 7.995958 | 7.808132 | 5.570536 | 2.854199 |
| 50% | 14.510081 | 9.080156 | 8.493766 | 6.487903 | 3.216644 |
| 75% | 15.718750 | 9.911365 | 9.595357 | 7.602959 | 3.754963 |
| max | 19.276087 | 12.905914 | 11.522098 | 9.114247 | 6.542719 |

| | LONG BEACH | ALAMEDA | BAYTOWN | SAVANNAH | MIAMI |
|---|---|---|---|---|---|
| count | 82.000000 | 82.000000 | 82.000000 | 82.000000 | 82.000000 |
| mean | 7.076181 | 8.747704 | 4.240066 | 7.139021 | 5.464240 |
| std | 1.546099 | 1.761082 | 0.572711 | 1.008751 | 1.241860 |
| min | 4.474326 | 6.126344 | 3.053314 | 5.108333 | 2.957910 |
| 25% | 6.163686 | 7.632056 | 3.816113 | 6.463539 | 4.482992 |
| 50% | 6.635729 | 8.418118 | 4.226478 | 7.061156 | 5.410205 |
| 75% | 7.663371 | 9.490062 | 4.518262 | 7.663812 | 6.152767 |
| max | 11.760030 | 14.509722 | 6.029983 | 10.285297 | 8.343750 |

| | TIPS1-5 | TIPS5 | TIPS10 | Oil | ISMM | NFPTOT | Gd/Co |
|---|---|---|---|---|---|---|---|
| count | 82.00 | 82.00 | 82.00 | 82.00 | 82.00 | 82.00 | 82.00 |
| mean | 144.11 | -0.06 | 0.21 | 56.47 | 55.01 | 146172.17 | -0.06 |
| std | 8.47 | 0.82 | 0.63 | 11.80 | 4.78 | 4075.14 | 1.12 |
| min | 135.05 | -1.85 | -1.08 | 33.09 | 41.50 | 130513.00 | -2.29 |
| 25% | 139.28 | -0.10 | 0.04 | 47.86 | 51.35 | 143672.25 | -0.84 |
| 50% | 140.37 | 0.12 | 0.43 | 55.96 | 55.35 | 146486.50 | -0.05 |
| 75% | 148.41 | 0.39 | 0.65 | 65.73 | 59.03 | 149311.75 | 0.94 |
| max | 165.25 | 1.00 | 0.98 | 81.39 | 64.70 | 152504.00 | 2.33 |

| | InflationSwap5Y5Y | BED | PERFSP500 | SLOPE2_10 | M2Growth | CRB |
|---|---|---|---|---|---|---|
| count | 82.00 | 82.00 | 82.00 | 82.00 | 82.00 | 82.00 |
| mean | 2.25 | 885.68 | 1.03 | 0.76 | 9.16 | 435.30 |
| std | 0.17 | 273.78 | 0.07 | 0.44 | 6.70 | 58.04 |
| min | 1.81 | 466.00 | 0.76 | -0.01 | 3.20 | 353.23 |
| 25% | 2.11 | 684.25 | 1.00 | 0.36 | 4.95 | 402.66 |
| 50% | 2.26 | 858.00 | 1.04 | 0.79 | 6.30 | 419.22 |
| 75% | 2.38 | 1016.75 | 1.07 | 1.08 | 12.18 | 440.42 |
| max | 2.61 | 1597.00 | 1.26 | 1.73 | 26.90 | 615.16 |

# C Summary Statistics (Daily Data)

| | NORSWORTHY | PORT EVERGLADES | BAYONNE | GRETNA | GALVESTON |
|---|---|---|---|---|---|
| count | 1627.00 | 1627.00 | 1627.00 | 1627.00 | 1627.00 |
| mean | 14.62 | 9.15 | 8.39 | 6.48 | 3.38 |
| std | 3.78 | 2.77 | 2.49 | 2.19 | 1.52 |
| min | 2.62 | 1.00 | 1.33 | 1.00 | 1.00 |
| 25% | 12.08 | 7.19 | 6.58 | 4.92 | 2.25 |
| 50% | 14.75 | 9.08 | 8.38 | 6.38 | 3.21 |
| 75% | 17.12 | 10.88 | 10.00 | 7.88 | 4.25 |
| max | 26.83 | 18.83 | 17.33 | 14.38 | 10.08 |

| | LONG BEACH | ALAMEDA | BAYTOWN | SAVANNAH | MIAMI |
|---|---|---|---|---|---|
| count | 1627.00 | 1627.00 | 1627.00 | 1627.00 | 1627.00 |
| mean | 7.23 | 8.67 | 4.25 | 7.16 | 5.52 |
| std | 2.31 | 2.33 | 1.53 | 2.09 | 2.00 |
| min | 1.54 | 2.38 | 1.00 | 1.00 | 1.00 |
| 25% | 5.58 | 7.08 | 3.12 | 5.71 | 4.08 |
| 50% | 7.12 | 8.42 | 4.17 | 7.04 | 5.38 |
| 75% | 8.83 | 9.98 | 5.21 | 8.50 | 6.79 |
| max | 15.92 | 17.96 | 9.62 | 14.60 | 12.50 |

| | TIPS1-5 | TIPS5 | TIPS10 | Oil | ISMM | NFPTOT | Gd/Co |
|---|---|---|---|---|---|---|---|
| count | 1627.00 | 1627.00 | 1627.00 | 1627.00 | 1627.00 | 1627.00 | 1627.00 |
| mean | 145.38 | -0.19 | 0.11 | 57.52 | 54.87 | 145962.94 | -0.06 |
| std | 8.95 | 0.83 | 0.65 | 13.95 | 4.61 | 3974.49 | 1.08 |
| min | 134.63 | -1.91 | -1.19 | 18.44 | 40.96 | 129946.22 | -2.38 |
| 25% | 139.38 | -0.55 | -0.50 | 48.30 | 51.32 | 143172.33 | -0.85 |
| 50% | 140.89 | 0.06 | 0.33 | 58.31 | 55.09 | 146174.83 | -0.04 |
| 75% | 150.51 | 0.35 | 0.58 | 67.16 | 58.99 | 149184.77 | 0.84 |
| max | 166.94 | 1.16 | 1.15 | 116.29 | 63.81 | 153519.04 | 2.63 |

| | InflationSwap5Y5Y | BED | PERFSP500 | SLOPE2_10 | M2Growth | CRB |
|---|---|---|---|---|---|---|
| count | 1627.00 | 1627.00 | 1627.00 | 1627.00 | 1627.00 | 1627.00 |
| mean | 2.22 | 770.49 | 1.03 | 0.78 | 9.02 | 434.51 |
| std | 0.21 | 220.28 | 0.07 | 0.46 | 6.57 | 56.28 |
| min | 1.22 | 403.00 | 0.69 | -0.05 | 3.18 | 347.55 |
| 25% | 2.07 | 637.00 | 1.00 | 0.35 | 4.96 | 404.14 |
| 50% | 2.25 | 742.00 | 1.04 | 0.80 | 6.18 | 417.36 |
| 75% | 2.39 | 847.50 | 1.07 | 1.16 | 12.08 | 439.54 |
| max | 2.70 | 1958.00 | 1.39 | 1.78 | 26.72 | 632.94 |

# D Raw average number of ships at ports v MA of 22 days

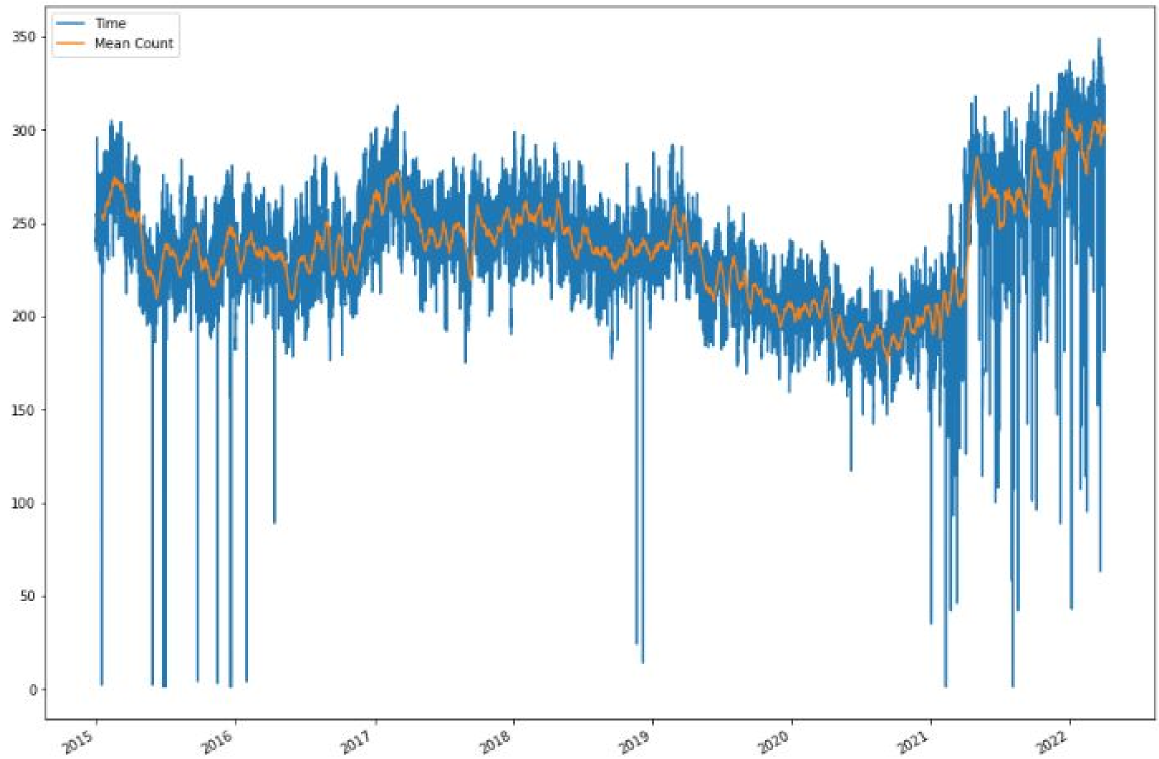


Figure 24: Raw Number of ships v Cleaned Moving Average.

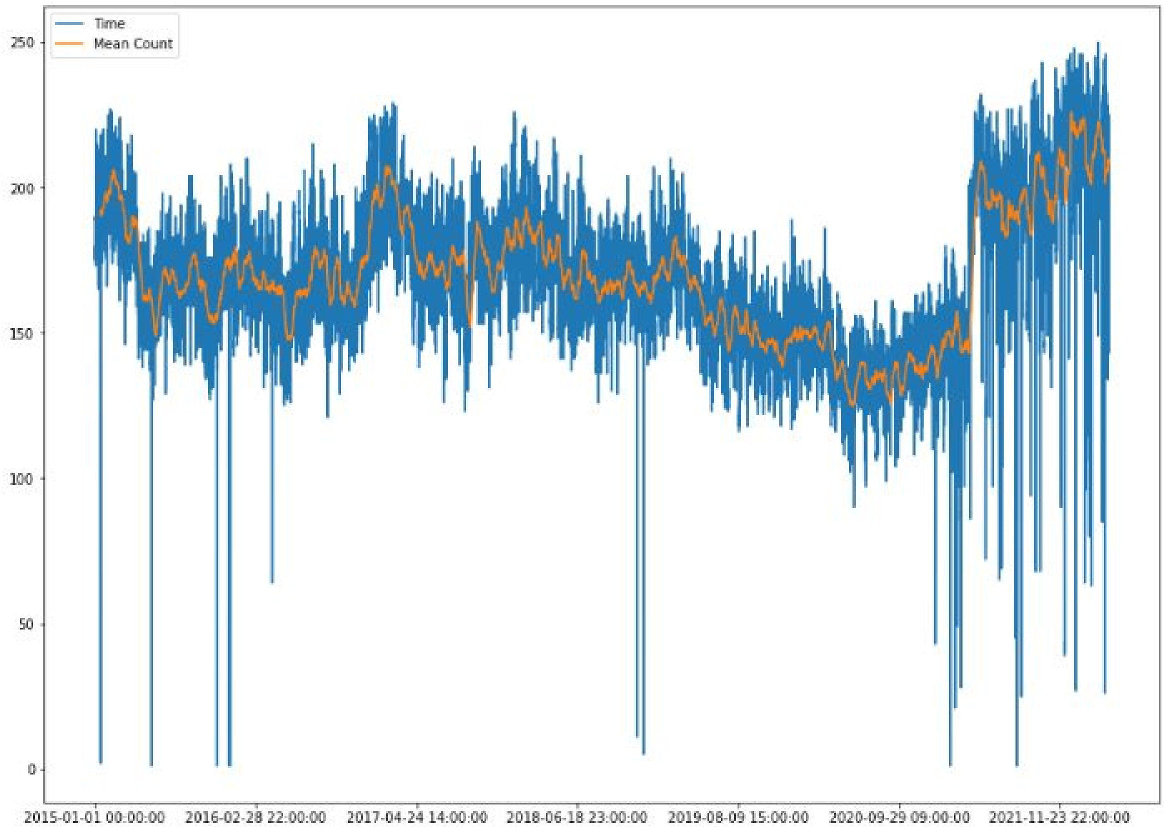


Figure 25: Raw Number of ships in the top 40 ports v Cleaned Moving Average.

# E Best and Worst Performing Model Combinations

| **Features** | **MSE** |
|---|---|
| CPI Diff LAG, Port Time, PORT EVERGLADES, BAYTOWN | 0.072 |
| CPI Diff LAG, PORT EVERGLADES, BAYTOWN, MIAMI | 0.073 |
| PORT EVERGLADES, BAYTOWN, MIAMI, CPI Diff LAG | 0.073 |
| PORT EVERGLADES, GRETNA, ALAMEDA, BAYTOWN, CPI Diff LAG | 0.073 |
| PORT EVERGLADES, GRETNA, BAYTOWN, MIAMI, CPI Diff LAG | 0.075 |
| PORT EVERGLADES, ALAMEDA, BAYTOWN, CPI Diff LAG | 0.075 |
| BAYONNE, LONG BEACH, ALAMEDA, SAVANNAH, MIAMI, CPI Diff LAG | 0.076 |
| Port Time, PORT EVERGLADES, ALAMEDA, BAYTOWN, CPI Diff LAG | 0.076 |
| BAYONNE, LONG BEACH, ALAMEDA, MIAMI, CPI Diff LAG | 0.079 |
| CPI Diff LAG, PORT EVERGLADES, BAYTOWN | 0.079 |

Table 17: Top 10 feature combinations with Lowest MSE Monthly using Monthly Shipping Data.

| **Features** | **MSE** |
|---|---|
| Mean Number, Port Time, NORSWORTHY, PORT EVERGLADES, GRETNA, SAVANNAH, MIAMI, CPI Diff LAG | 0.237 |
| Mean Number, Mean Top40, Port Time, NORSWORTHY, PORT EVERGLADES, GALVESTON, CPI Diff LAG | 0.237 |
| Mean Number, Port Time, NORSWORTHY, PORT EVERGLADES, GALVESTON, SAVANNAH, CPI Diff LAG | 0.238 |
| Mean Top40, Port Time, NORSWORTHY, PORT EVERGLADES, GALVESTON, ALAMEDA, SAVANNAH, CPI Diff LAG | 0.239 |
| Mean Number, Port Time, NORSWORTHY, PORT EVERGLADES, GRETNA, GALVESTON, ALAMEDA, SAVANNAH, CPI Diff LAG | 0.240 |
| Port Time, NORSWORTHY, CPI Diff LAG | 0.241 |
| Port Time, NORSWORTHY, SAVANNAH, CPI Diff LAG | 0.243 |
| Mean Top40, Port Time, NORSWORTHY, PORT EVERGLADES, BAYONNE, GALVESTON, SAVANNAH, CPI Diff LAG | 0.244 |
| Mean Top40, Port Time, NORSWORTHY, PORT EVERGLADES, BAYONNE, ALAMEDA, SAVANNAH, CPI Diff LAG | 0.244 |
| Mean Number, Mean Top40, Port Time, NORSWORTHY, PORT EVERGLADES, BAYONNE, GALVESTON, ALAMEDA, SAVANNAH, CPI Diff LAG | 0.247 |

Table 18: Bottom 10 feature combinations with Highest MSE using Monthly Shipping Data.

| **Features** | **MSE** |
|---|---|
| Mean Top40, BAYONNE(T), ALAMEDA(T), PORT EVERGLADES(T), CPI Lag | 0.062 |
| Mean Top40, ALAMEDA(T), PORT EVERGLADES(T), CPI Lag | 0.071 |
| Mean Top40, NORSWORTHY(T), BAYONNE(T), ALAMEDA(T), PORT EVERGLADES(T), CPI Lag | 0.073 |
| Port Time, SAVANNAH(T), CPI Lag | 0.076 |
| Mean Top40, BAYONNE(T), ALAMEDA(T), MIAMI(T), PORT EVERGLADES(T), CPI Lag | 0.076 |
| Mean Top40, BAYONNE(T), ALAMEDA(T), SAVANNAH(T), PORT EVERGLADES(T), CPI Lag | 0.077 |
| Mean Top40, BAYONNE(T), ALAMEDA(T), SAVANNAH(T), MIAMI(T), PORT EVERGLADES(T), CPI Lag | 0.077 |
| Port Time, SAVANNAH(T), PORT EVERGLADES(T), CPI Lag | 0.078 |
| Mean Top40, NORSWORTHY(T), BAYONNE(T), ALAMEDA(T), SAVANNAH(T), PORT EVERGLADES(T), CPI Lag | 0.078 |
| Port Time, GRETNA(T), SAVANNAH(T), PORT EVERGLADES(T), CPI Lag | 0.079 |

Table 19: Top 10 feature combinations with Lowest MSE using Daily Shipping Data.

| **Features** | **MSE** |
|---|---|
| NORSWORTHY(T), BAYONNE(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.283 |
| Port Time, NORSWORTHY(T), LONG BEACH(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.290 |
| NORSWORTHY(T), LONG BEACH(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.293 |
| Port Time, BAYONNE(T), LONG BEACH(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.293 |
| ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.298 |
| Port Time, NORSWORTHY(T), BAYONNE(T), LONG BEACH(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.300 |
| Port Time, NORSWORTHY(T), BAYONNE(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.301 |
| BAYONNE(T), LONG BEACH(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.309 |
| Port Time, BAYONNE(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.321 |
| BAYONNE(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.341 |

Table 20: Bottom 10 feature combinations with Highest MSE using Daily Shipping Data.

| **Features** | **MSE** |
|---|---|
| Mean Top40, BAYONNE(T), ALAMEDA(T), PORT EVERGLADES(T), CPI Lag | 0.062 |
| Mean Top40, ALAMEDA(T), PORT EVERGLADES(T), CPI Lag | 0.071 |
| Mean Top40, NORSWORTHY(T), BAYONNE(T), ALAMEDA(T), PORT EVERGLADES(T), CPI Lag | 0.073 |
| Port Time, SAVANNAH(T), CPI Lag | 0.076 |
| Mean Top40, BAYONNE(T), ALAMEDA(T), MIAMI(T), PORT EVERGLADES(T), CPI Lag | 0.076 |
| Mean Top40, BAYONNE(T), ALAMEDA(T), SAVANNAH(T), PORT EVERGLADES(T), CPI Lag | 0.077 |
| Mean Top40, BAYONNE(T), ALAMEDA(T), SAVANNAH(T), MIAMI(T), PORT EVERGLADES(T), CPI Lag | 0.077 |
| Port Time, SAVANNAH(T), PORT EVERGLADES(T), CPI Lag | 0.078 |
| Mean Top40, NORSWORTHY(T), BAYONNE(T), ALAMEDA(T), SAVANNAH(T), PORT EVERGLADES(T), CPI Lag | 0.078 |
| Port Time, GRETNA(T), SAVANNAH(T), PORT EVERGLADES(T), CPI Lag | 0.079 |

Table 21: Top 10 feature combinations with Lowest MSE using Daily Shipping Data.

| **Features** | **MSE** |
|---|---|
| NORSWORTHY(T), BAYONNE(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.283 |
| Port Time, NORSWORTHY(T), LONG BEACH(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.290 |
| NORSWORTHY(T), LONG BEACH(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.293 |
| Port Time, BAYONNE(T), LONG BEACH(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.293 |
| ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.298 |
| Port Time, NORSWORTHY(T), BAYONNE(T), LONG BEACH(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.300 |
| Port Time, NORSWORTHY(T), BAYONNE(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.301 |
| BAYONNE(T), LONG BEACH(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.309 |
| Port Time, BAYONNE(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.321 |
| BAYONNE(T), ALAMEDA(T), SAVANNAH(T), CPI Lag | 0.341 |

Table 22: Bottom 10 feature combinations with Highest MSE using Daily Shipping Data.

| **Features** | **MSE** |
|---|---|
| BAYONNE(T), Mean Top40, ALAMEDA(T), PORT EVERGLADES(T), CPI Lag | 0.062 |
| BAYONNE(T), Mean Top40, TIPS10, ALAMEDA(T), MIAMI(T), PORT EVERGLADES(T), CPI Lag | 0.070 |
| Mean Top40, ALAMEDA(T), PORT EVERGLADES(T), CPI Lag | 0.071 |
| BAYONNE(T), Mean Top40, TIPS10, ALAMEDA(T), PORT EVERGLADES(T), CPI Lag | 0.072 |
| BAYONNE(T), Mean Top40, ALAMEDA(T), InflationSwap5Y5Y, PORT EVERGLADES(T), CPI Lag | 0.076 |
| BAYONNE(T), Mean Top40, ALAMEDA(T), MIAMI(T), PORT EVERGLADES(T), CPI Lag | 0.077 |
| BAYONNE(T), Mean Top40, TIPS10, ALAMEDA(T), MIAMI(T), Oil, PORT EVERGLADES(T), CPI Lag | 0.080 |
| BAYONNE(T), Mean Top40, TIPS10, ALAMEDA(T), Oil, PORT EVERGLADES(T), CPI Lag | 0.081 |
| BAYONNE(T), Mean Top40, TIPS10, ALAMEDA(T), MIAMI(T), CPI Lag | 0.082 |
| BAYONNE(T), Mean Top40, ALAMEDA(T), CPI Lag | 0.083 |

Table 23: Top 10 feature combinations with Lowest MSE using Daily Shipping and Market Data.

| Features | MSE |
|---|---|
| BAYONNE(T), Mean Top40, TIPS10, Oil, GOLDCOPPER, InflationSwap5Y5Y, PORT EVERGLADES(T), CPI Lag | 0.395 |
| BAYONNE(T), Mean Top40, GOLDCOPPER, PORT EVERGLADES(T), CPI Lag | 0.398 |
| BAYONNE(T), Mean Top40, GOLDCOPPER, InflationSwap5Y5Y, PORT EVERGLADES(T), CPI Lag | 0.404 |
| Oil, CPI Lag | 0.404 |
| TIPS10, Oil, CPI Lag | 0.405 |
| BAYONNE(T), Mean Top40, TIPS10, GOLDCOPPER, InflationSwap5Y5Y, PORT EVERGLADES(T), CPI Lag | 0.408 |
| Mean Top40, GOLDCOPPER, PORT EVERGLADES(T), CPI Lag | 0.411 |
| Mean Top40, TIPS10, GOLDCOPPER, PORT EVERGLADES(T), CPI Lag | 0.415 |
| Mean Top40, TIPS10, GOLDCOPPER, InflationSwap5Y5Y, PORT EVERGLADES(T), CPI Lag | 0.422 |
| Mean Top40, GOLDCOPPER, InflationSwap5Y5Y, PORT EVERGLADES(T), CPI Lag | 0.427 |

Table 24: Botoom 10 feature combinations with Highest MSE using Daily Shipping and Market Data.

# F AR and VAR (Benchmark) Result Summaries (Monthly and Daily)

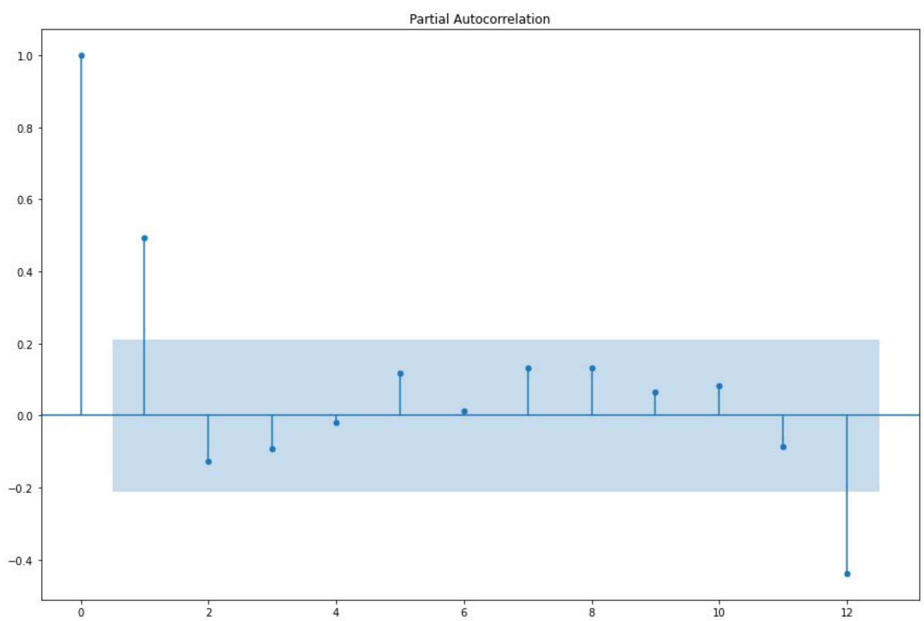


Figure 26: Partial Autocorrelation Results - Monthly Data.

```
Regression Statistics
                            AutoReg Model Results
==============================================================================
Dep. Variable:               CPI_diff   No. Observations:                   75
Model:                     AutoReg(2)   Log Likelihood                 -17.129
Method:               Conditional MLE   S.D. of innovations              0.306
Date:                Sat, 02 Jul 2022   AIC                             -2.259
Time:                        13:37:11   BIC                             -2.133
Sample:                    05-01-2015   HQIC                            -2.209
                         - 05-01-2021
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
intercept       0.0418      0.036      1.159      0.246      -0.029       0.113
CPI_diff.L1     0.6483      0.126      5.148      0.000       0.401       0.895
CPI_diff.L2    -0.2440      0.133     -1.830      0.067      -0.505       0.017
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.3283           -1.5276j            2.0243           -0.1361
AR.2            1.3283           +1.5276j            2.0243            0.1361
-----------------------------------------------------------------------------
```

Figure 27: AR Result Summaries - Monthly Data.

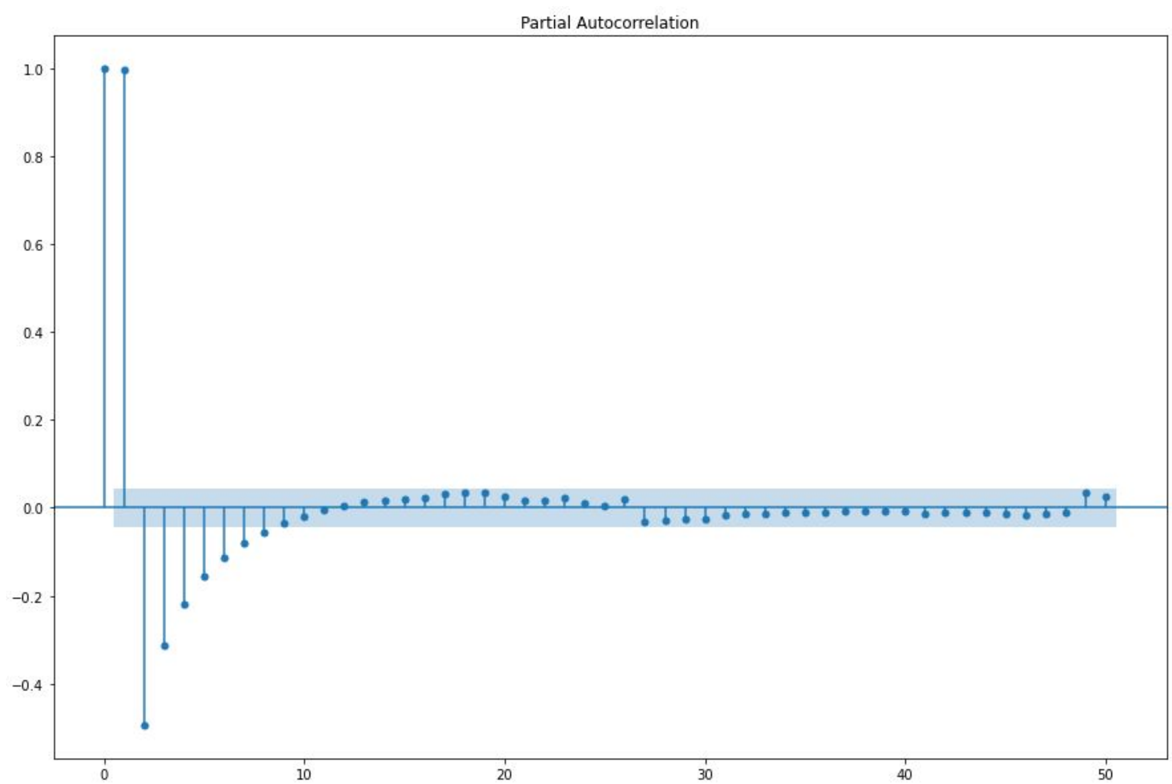


Figure 28: Partial Autocorrelation Results - Daily Data.

```
Regression Statistics
                            AutoReg Model Results
==============================================================================
Dep. Variable:               CPI_diff   No. Observations:                 1639
Model:                    AutoReg(10)   Log Likelihood                6809.113
Method:               Conditional MLE   S.D. of innovations              0.004
Date:                Thu, 30 Jun 2022   AIC                            -11.183
Time:                        11:08:44   BIC                            -11.143
Sample:                    02-17-2015   HQIC                           -11.168
                         - 05-14-2021
================================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
intercept       0.0001   9.28e-05      1.174      0.240   -7.29e-05       0.000
CPI_diff.L1     1.8843      0.025     76.247      0.000       1.836       1.933
CPI_diff.L2    -0.7507      0.053    -14.218      0.000      -0.854      -0.647
CPI_diff.L3    -0.1348      0.056     -2.408      0.016      -0.244      -0.025
CPI_diff.L4    -0.0115      0.056     -0.205      0.838      -0.121       0.098
CPI_diff.L5     0.0010      0.056      0.017      0.986      -0.109       0.111
CPI_diff.L6     0.0092      0.056      0.164      0.870      -0.101       0.119
CPI_diff.L7     0.0404      0.056      0.720      0.471      -0.070       0.150
CPI_diff.L8    -0.0459      0.056     -0.816      0.415      -0.156       0.064
CPI_diff.L9    -0.0655      0.053     -1.231      0.218      -0.170       0.039
CPI_diff.L10    0.0718      0.025      2.880      0.004       0.023       0.121
                                    Roots
==============================================================================
                  Real          Imaginary           Modulus         Frequency
------------------------------------------------------------------------------
AR.1           -1.5074           -0.0000j            1.5074           -0.5000
AR.2           -1.1192           -0.9338j            1.4576           -0.3893
AR.3           -1.1192           +0.9338j            1.4576            0.3893
AR.4           -0.1540           -1.3818j            1.3903           -0.2677
AR.5           -0.1540           +1.3818j            1.3903            0.2677
AR.6            0.8765           -1.0334j            1.3551           -0.1380
AR.7            0.8765           +1.0334j            1.3551            0.1380
AR.8            1.0288           -0.0480j            1.0299           -0.0074
AR.9            1.0288           +0.0480j            1.0299            0.0074
AR.10           1.1547           -0.0000j            1.1547           -0.0000
------------------------------------------------------------------------------
```

Figure 29: AR Result Summaries - Daily Data.

```
  Summary of Regression Results   
==================================
Model:                         VAR
Method:                        OLS
Date:           Mon, 08, Aug, 2022
Time:                     12:54:52
--------------------------------------------------------------------
No. of Equations:         6.00000    BIC:                    15.4097
Nobs:                     74.0000    HQIC:                   13.9499
Log likelihood:          -1032.31    FPE:                    443819.
AIC:                      12.9811    Det(Omega_mle):         168066.
--------------------------------------------------------------------
Results for equation CPI
=====================================================================================
                        coefficient       std. error           t-stat            prob
-------------------------------------------------------------------------------------
const                      4.326425         2.727938            1.586           0.113
L1.CPI                     0.513886         0.158771            3.237           0.001
L1.NFPTOT                 -0.000028         0.000022           -1.299           0.194
L1.ISMM                    0.002056         0.019832            0.104           0.917
L1.US6M_MA                 0.104155         0.703742            0.148           0.882
L1.port_time_clean         0.000324         0.025668            0.013           0.990
L1.Mean Top40              0.009514         0.005431            1.752           0.080
L2.CPI                    -0.087038         0.146174           -0.595           0.552
L2.NFPTOT                 -0.000004         0.000022           -0.172           0.863
L2.ISMM                    0.005380         0.021472            0.251           0.802
L2.US6M_MA                -0.046824         0.689939           -0.068           0.946
L2.port_time_clean        -0.013080         0.024177           -0.541           0.589
L2.Mean Top40             -0.009950         0.005105           -1.949           0.051
=====================================================================================
```

Figure 30: VAR Bench Summary results for the CPI variable - Monthly Data.

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Mon, 08, Aug, 2022
Time:                     12:55:00
--------------------------------------------------------------------
No. of Equations:         6.00000    BIC:                   -30.1086
Nobs:                     1681.00    HQIC:                  -30.2672
Log likelihood:           11284.5    FPE:                6.52575e-14
AIC:                     -30.3604    Det(Omega_mle):     6.23098e-14
--------------------------------------------------------------------
Results for equation CPI
=====================================================================================
                         coefficient       std. error           t-stat            prob
-------------------------------------------------------------------------------------
const                       0.001724         0.007196            0.240           0.811
L1.US6M_MA                 -0.022223         0.036401           -0.610           0.542
L1.port_time_clean          0.000064         0.000259            0.247           0.805
L1.Mean Top40               0.000062         0.000062            1.004           0.315
L1.CPI                      1.961294         0.006554          299.243           0.000
L1.NFPTOT                   0.000003         0.000001            2.335           0.020
L1.ISMM                    -0.001913         0.001184           -1.615           0.106
L2.US6M_MA                  0.021961         0.036443            0.603           0.547
L2.port_time_clean         -0.000047         0.000258           -0.182           0.855
L2.Mean Top40              -0.000054         0.000062           -0.869           0.385
L2.CPI                     -0.964366         0.006562         -146.953           0.000
L2.NFPTOT                  -0.000003         0.000001           -2.368           0.018
L2.ISMM                     0.001959         0.001189            1.648           0.099
=====================================================================================
```

Figure 31: VAR Bench Summary results for the CPI variable - Daily Data.

# G Counterfactual Inflation Impact Estimation - OLS Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | CPI | **R-squared:** | 0.314 |
| **Model:** | OLS | **Adj. R-squared:** | 0.297 |
| **Method:** | Least Squares | **F-statistic:** | 18.97 |
| **Date:** | Thu, 15 Sep 2022 | **Prob (F-statistic):** | 1.64e-07 |
| **Time:** | 09:55:10 | **Log-Likelihood:** | -18.796 |
| **No. Observations:** | 86 | **AIC:** | 43.59 |
| **Df Residuals:** | 83 | **BIC:** | 50.95 |
| **Df Model:** | 2 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P>\|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **const** | -0.6480 | 0.293 | -2.214 | 0.030 | -1.230 | -0.066 |
| **CPI_lag** | 0.4151 | 0.095 | 4.385 | 0.000 | 0.227 | 0.603 |
| **Mean Top40** | 0.0042 | 0.002 | 2.431 | 0.017 | 0.001 | 0.008 |

| | | | |
|---|---|---|---|
| **Omnibus:** | 4.972 | **Durbin-Watson:** | 1.663 |
| **Prob(Omnibus):** | 0.083 | **Jarque-Bera (JB):** | 5.575 |
| **Skew:** | 0.252 | **Prob(JB):** | 0.0616 |
| **Kurtosis:** | 4.141 | **Cond. No.** | 1.52e+03 |

Figure 32: OLS regression summary for the counterfactual study.