Predicting with Python

JTBQ-025
The Anatomies of Economic Crises
PD Dr. Christian Müller
Constructor University

Stephan Daaboul
Jan Steinmüller
Oscar Echeverria

Part 1 Recession Prediction

Goals and Objectives

In this study, we sought to investigate whether it is possible to predict recessions and, if so, which type of model might be best suited for this task. We also examined the challenges of working with economic data and identified the most significant indicators for predicting recessions. Our goal was to determine the feasibility of reliably predicting future recessions in a real-world scenario. Specifically, we aimed to answer the following questions: (1) Can we predict recessions using machine learning models? (2) Which model performs best for this task? (3) What are the challenges of working with economic data and how do they impact our predictions? (4) Which indicators are the most significant for predicting recessions? (5) Is it feasible to reliably predict future recessions in a real-world scenario?

Data and Assumptions

To predict future recessions, we need to make a few assumptions about the nature of recessions and how they can be predicted. The first assumption is that recessions are, at least to some extent, predictable. This means that there are certain patterns or indicators that can signal the onset of a recession. The second assumption is that these indicators are clear and discernible, meaning that we can identify them through data analysis or other means. The third assumption is that future recessions will have similar characteristics to past recessions, in terms of their causes, timing, and effects. This means that we can draw upon our understanding of historical recessions to inform our predictions of future recessions.

These assumptions are not necessarily true in all cases, as recessions can be driven by unexpected events or factors. However, they provide a starting point for making predictions and allow us to test the accuracy and reliability of different models and approaches. By carefully evaluating these assumptions and the underlying data, we can improve our understanding of recessions and potentially develop more accurate methods for predicting them.

The data that we are analyzing in this study comes from multiple sources, with the majority coming from the Federal Reserve Economic Data (FRED). From FRED, we obtained various economic indicators such as GDP, unemployment rate, housing market index, and annual interest rates on 10-year and 3-month US treasury bonds. We also obtained more qualitative data from the University of Michigan, which conducts surveys on the likelihood of citizens buying large household items (such as fridges and washing machines) in the current quarter. These survey results are represented as the 'good time to buy' and 'bad time to buy' features. After preprocessing the data, we ended up with a total of 28 different features, which we will later examine in terms of their importance and significance.

Theoretical background

In order to predict recessions, we are using three different approaches: logistic regression, random forests, and XGBoost. We will compare the performance of these approaches and examine the features they prioritize in their predictions. By comparing these methods' performance and feature importance, we aim to identify which approach is most effective for predicting recessions and which economic indicators are most relevant for this task.

It is worth noting that all of the approaches we will be using in this study are binary classifiers, meaning that they will only predict whether a recession will occur or not. While these models may provide valuable insights into the likelihood of a recession occurring, they do not provide any additional information about the potential length or severity of a recession.

The first approach that we will be using is logistic regression. Logistic regression is a statistical method used for classification tasks, where the goal is to predict a binary outcome (e.g., recession or no recession). It is based on the idea of using a linear model to predict the probability that an event will occur (e.g., the probability that a recession will occur). The model is trained using labelled data, and the output is a probability between 0 and 1 that the event will occur. The probability is then converted into a binary prediction using a threshold value (e.g., if the probability is greater than 0.5, the prediction is "recession," otherwise it is "no recession"). This approach is also the most trivial one of the three. However, it is true that logistic regression can be sensitive to noisy data and can be prone to overfitting, particularly when the dataset is small or imbalanced. Overfitting occurs when a model is too complex and fits the training data too well, leading to a poor generalization of new data. This can be a concern in the context of recession prediction, as the goal is to build a model that can accurately predict recessions in the future, not just fit the training data well.

The next method we use is random forests. Random forest regression involves generating a set of decision trees in parallel, where each tree has a different structure (e.g., number of branches, nodes, and size). To make a prediction, the model takes the average of the predictions made by all the trees and classifies the data point based on this average. The use of multiple trees with different structures and the averaging of their predictions helps to reduce the risk of overfitting and makes random forests effective at predicting data that is difficult to predict, such as imbalanced or small datasets. This makes random forests a popular choice for financial and economic predictions, where data can often be imbalanced and scarce. Random forest also converges faster than logistic regression because of its parallel generation of trees and its randomness being more resistant to noisy data like ours.

The final method we will be using is XGBoost, which stands for eXtreme Gradient Boosting. Like random forests, XGBoost uses decision trees to make predictions. However, the trees in XGBoost are generated sequentially and the gradient of each current tree is taken into account when generating the next tree. This allows XGBoost to converge more quickly than

random forests, as it does not have to generate new trees from scratch that might have poor performance.
One potential downside of using XGBoost is that it may be more prone to getting stuck in local minima, as it is based on the gradient of the current trees. This can make it less effective at predicting hard-to-predict data, such as imbalanced or small datasets. However, XGBoost can be a powerful tool when the data is well-behaved and there is a clear gradient to follow.

Results of Convergence

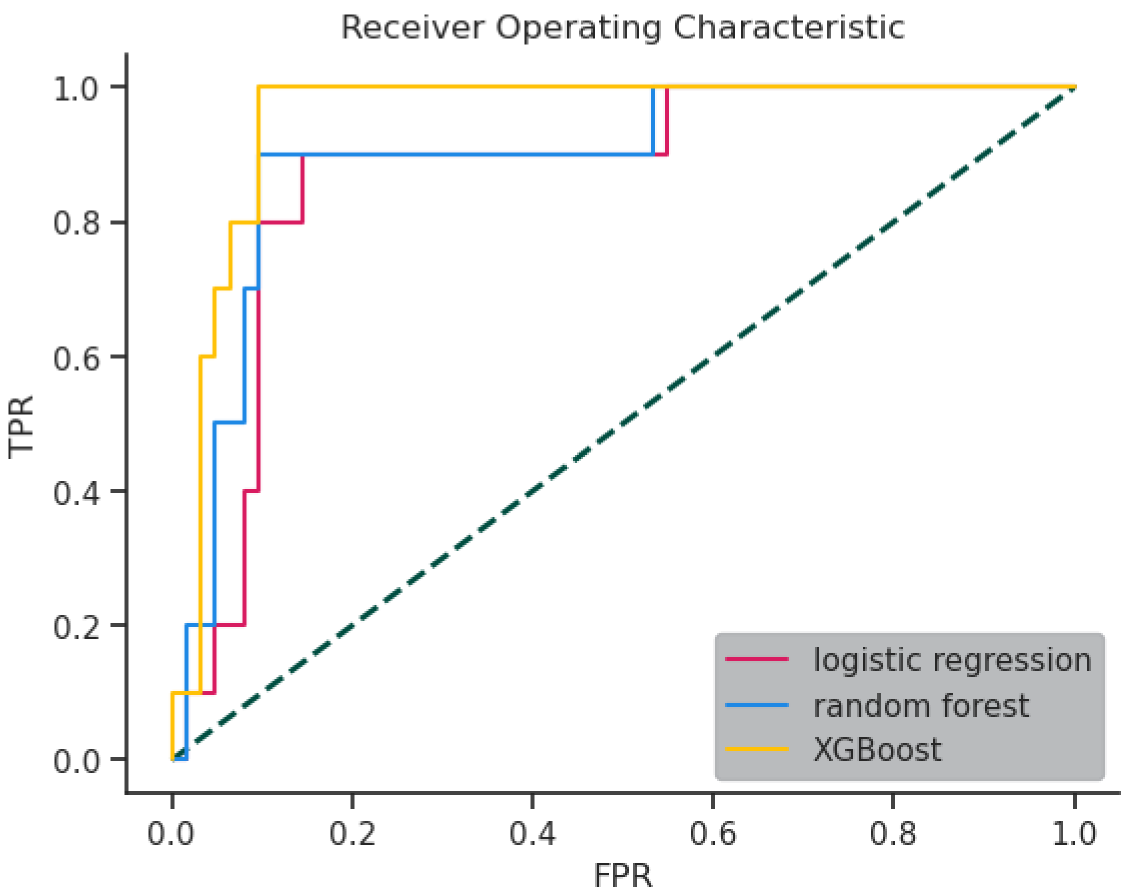


Figure 1.1

When running our models we got the following AUC (Area Under the Curve):

```
Area under the curve logistic regression: 0.8694
Area under the curve random forest: 0.8935
Area under the curve XGBoost: 0.9532
```

From the AUC and the diagram we can see that XGBoost converges the fastest since it is using the gradient in order to generate the next tree. Logistic regression is performing and converging almost the same as random forest. However, we might have overfitted our training data which could make it hard to predict future recessions. Our random forest model is more robust against that.

In summary, logistic regression is the simplest approach among the three methods, but it can be prone to overfitting, particularly when the dataset is small or imbalanced. XGBoost is the fastest model, but it may be more susceptible to getting stuck in local minima, which can impact its ability to accurately predict hard-to-predict data. Random forests are the most robust of the three methods, thanks to their ability to generate multiple decision trees with different structures and average their predictions. This makes random forests resistant to overfitting and well-suited for handling imbalanced and small datasets.

Accuracy of our models

Let us now evaluate our random forest model, as it should be the most robust. When evaluating the performance of our random forest model, it is important to focus on the recall rate, which measures the ability of the model to correctly identify true positives (i.e., predict a recession when one occurs). While the accuracy of the model is also important, it is less relevant in the context of recession prediction, as we are primarily interested in avoiding false negatives (i.e., not predicting a recession when one occurs). This is because the goal is to predict recessions as early as possible so that we can take appropriate action to mitigate their impact. On the other hand, predicting a recession when there is none (i.e., a false positive) may lead to unnecessary disruptions, but it is generally less harmful than failing to predict a recession. Therefore, it is essential to prioritize the recall rate when evaluating the performance of our random forest model for predicting recessions.

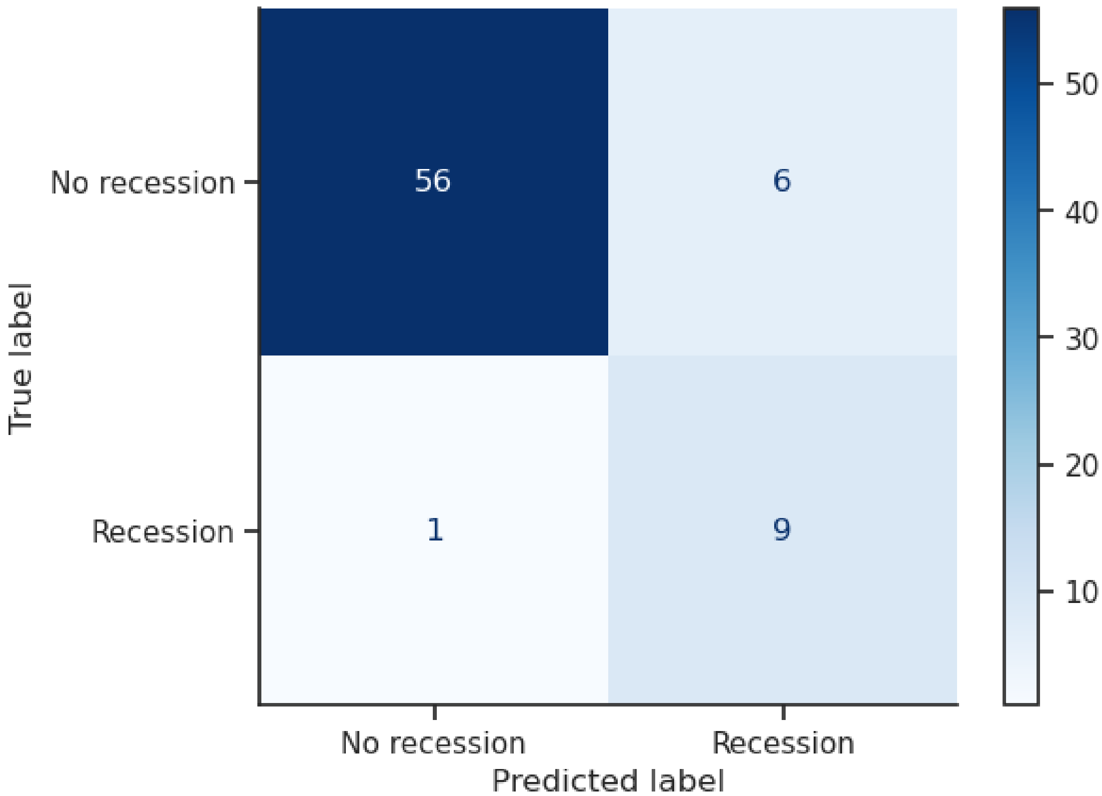


Figure 1.2

We can see that we have a recall rate of 90%. This is pretty good in terms of models. However, it also means that in 10 recessions we only predict 9. We can see this in our data as well. To get a better recall rate we would need more data (including more recessions) to train our model further. There might also be more indicators that could lead to a better prediction
In terms of actual time data Let’s now look at our actual time data. Here we can see when it actually predicted the recession.

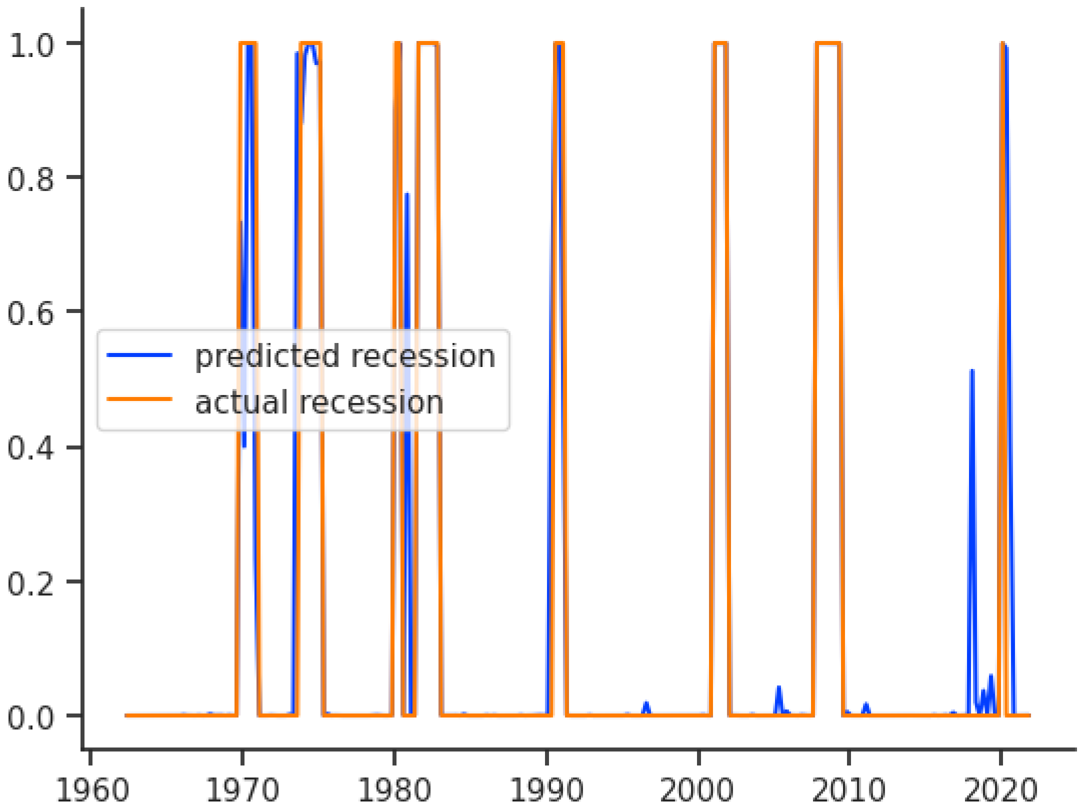


Figure 1.3

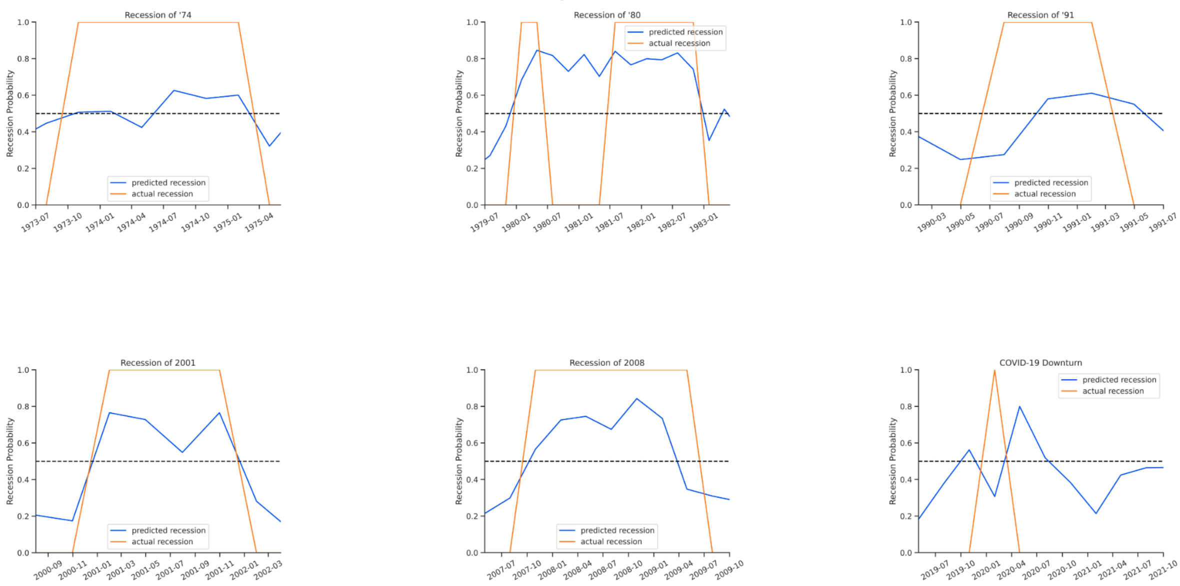


Figure 1.4

In the examples of historic recessions shown in the graph, the blue line represents the probability of a recession predicted by our model, while the orange line represents the actual occurrence of a recession. From the graph, it appears that our model is generally able to predict recessions accurately. However, there are some instances, such as the recession of 91 and the Covid-19 downturn, where the model may not have immediately recognized that a recession was occurring. This could be due to the lag between the onset of a recession and its impact on economic indicators, which our model relies on for prediction. As a result, it is difficult to determine with certainty whether our model actually predicted the recession or simply identified the effects of the recession after it had begun.

Most significant features

We will now be taking a look at which features are the most significant for each of the three models. To make the following graph more readable we will only be showing the ten most significant features.

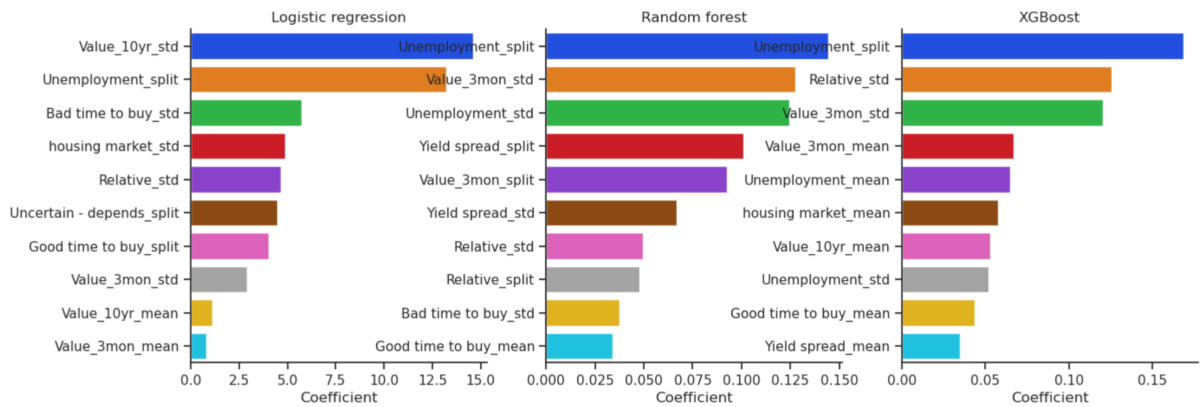


Figure 1.5

It was surprising to find that different features were more important for the different models. For logistic regression, we observed that the 'value10yr', which is the return of a 10-year treasury bond, and 'bad time to buy', which is a survey asking citizens if they are willing to buy large households items, features were among the top three most significant. However, these features did not rank as highly in random forests and XGBoost. We also noticed that the 'relative_std' and '3-month treasury bond values' features were more important to XGBoost compared to random forests. For random forests, we observed that the 'unemployment' and 'yield' features were more significant. Interestingly, the housing market was not among the top ten most important features for random forests.

From this analysis, we can conclude that the '3-month treasury bond' indicators contributed to the faster convergence of the models but may have also led them to a local minimum. The 'unemployment rate' appears to be the overall best indicator for predicting recessions. Additionally, we observed that the survey values 'good time to buy' and 'bad time to buy' were not particularly relevant in predicting recessions. Furthermore, the housing market did not seem to be as important as expected in predicting recessions.

Conclusion on the feasibility of real-time data

While we did identify some strong indicators of recessions in this analysis, some of the data, such as survey results on buying household items, may not be available in real-time. Additionally, we had to make some assumptions about recessions and the economy in order to make predictions, which may not always hold true in the real world. For example, recessions can be caused by factors that we did not consider in our analysis, as demonstrated by the Covid-19 downturn. In such cases, our models may not necessarily predict the recession itself, but rather detect the effects of the recession.

Another challenge in training these models is the limited amount of data on recessions, as there have not been many recessions in recent history. This can lead to imbalanced and noisy data, which can impact the performance of the models. Moreover, economies are constantly changing, which means that a model that works well at one point in time may not necessarily be applicable indefinitely. Therefore, it is important to regularly evaluate and update the models to ensure their continued effectiveness in predicting recessions.

Part 2 Consumer Price Index and Inflation Prediction

Introduction, Goals and Objectives

Consumer Price Index (CPI) measures the average price change over time for a basket of household goods and services. It is used to track changes in the cost of living and to adjust for inflation. The CPI is calculated by comparing the price of a basket of goods and services in a base period to the same basket in a current period and then expressing the result as an index number.

Inflation is the rate at which the general level of prices for goods and services is rising, and, subsequently, purchasing power is falling. Central banks attempt to limit inflation and avoid deflation to keep the economy running smoothly.

Inflation is typically measured by the percentage change in the CPI over a specific period, such as a month or a year. For example, if the CPI were 100 in January and 103 in February, the inflation rate for February would be 3%. This means that, on average, the prices of the goods and services in the basket have increased by 3% over February.
The main goal of this task is to predict the consumer price index (CPI) and inflation in an economy to anticipate changes in the cost of living and to provide policymakers and businesses with a sense of the direction in which prices are moving. This information can be used to make informed decisions about monetary and fiscal policy, business investments, and individual financial planning.

Several objectives may be pursued through the prediction of CPI and inflation:

1. Stabilizing prices: Inflation can disrupt economic activity and lead to uncertainty about the value of money. By predicting and controlling inflation, policymakers can stabilize prices and reduce the risk of deflation or hyperinflation.
2. Supporting economic growth: Accurate inflation predictions can help policymakers to set interest rates and other monetary policies that are conducive to economic growth.
3. Promoting financial stability: Inflation that is too high or too low can lead to financial instability. By predicting and controlling inflation, policymakers can promote financial stability.
4. Improving economic forecasting: Accurate inflation predictions can help businesses and individuals to make informed decisions about investments, saving, and spending. This can lead to the more efficient allocation of resources and support economic growth.
5. Facilitating comparisons: Inflation can make it challenging to compare prices over time. By predicting and controlling inflation, policymakers can facilitate comparisons and make it easier to compare prices across different periods.

Theoretical background

BOX PLOTS - Preparing Data

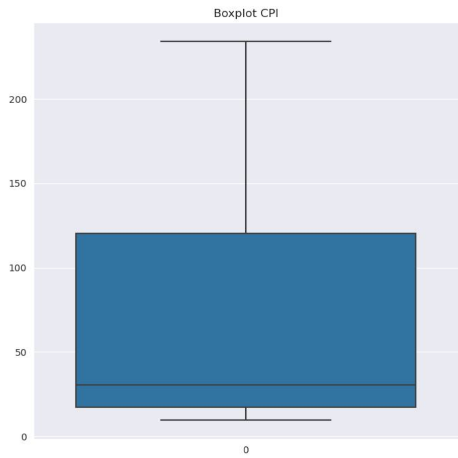
A box plot, also known as a box and whisker plot, is a graphical representation of a set of data that helps to summarise the distribution of the data and identify patterns and trends. It is a valuable tool for data visualization and statistical analysis, mainly when working with large datasets.

A box plot consists of a box representing the middle 50% of the data and several "whiskers" extending from the box to show the range of the data. The box is drawn from the lower quartile (25th percentile) to the upper quartile (75th percentile), with a line drawn at the median (50th percentile). The whiskers extend from the box to the minimum and maximum values in the data set unless the data contains outliers. The whiskers are drawn to the furthest data point within 1.5 times the interquartile range (the range between the lower and upper quartiles). *Outliers* are data points that are significantly different from the rest of the data and are plotted individually as dots.

Box plots help compare the distribution of data across different groups or conditions. For example, we could use a box plot to compare the distribution of test scores for two classes or height measurements for men and women. By looking at the box plots, we can see how the data is distributed and whether there are any significant differences between the groups.
Box plots are also useful for identifying outliers and potential errors in the data. Outliers can be caused by data entry errors, measurement errors, or other factors, and they can significantly impact the data analysis. By identifying and examining outliers, you can decide whether they should be included in the analysis or if they need to be corrected or removed.
Overall, box plots are a useful tool for understanding and summarising the distribution of a data set, identifying patterns and trends, and identifying potential outliers and errors in the data. As shown below in figure 2.1, it can be seen that the distribution of this data is between the acceptable –non-outlier– range. This means that the Consumer Price Index data follows

an average, acceptable trend that excludes any out-of-the-ordinary data caused by unpredictable economic changes that are not common and human error. The data is between the range of ~20 and ~125.

Figure 2.1



SEASONAL DECOMPOSITION - Better understanding the data

Seasonal decomposition is a statistical method that is commonly used in economics to analyze time series data, such as economic indicators like GDP, employment, inflation, and trade. By separating the data into its trend, seasonal, and residual components, it is possible to understand better the factors driving the data and identify any patterns or trends that may not be immediately apparent in the raw data.

The trend is the long-term direction of the data, which may be upward, downward, or flat. Seasonal patterns are repetitive patterns in the data that occur at regular intervals, such as monthly or annually. Residuals are random variations in the data that cannot be explained by the trend or seasonal components.

Additive seasonal decomposition can be helpful in understanding and interpreting time series data, mainly when the data exhibits a strong seasonal pattern. By separating the data into its trend, seasonal, and residual components, it is possible to understand better the factors driving the data and identify any patterns or trends that may not be immediately apparent in the raw data.

For example, suppose you are analyzing GDP data for a country. In that case, seasonal decomposition can help you understand whether the economy is growing or contracting over time and whether there are specific times of year (such as the holiday season) when economic activity tends to be higher or lower. It can also help you identify any unusual variations in the data (such as a sudden drop in GDP) that the trend or seasonal components may not explain.
Also, on a smaller scale, if you are analyzing sales data for a retail business, additive seasonal decomposition can help you understand whether the sales are increasing or decreasing over

time and whether there are any specific times of year (such as the holiday season) when sales tend to be higher or lower. It can also help you identify any unusual variations in the data (such as a sudden drop in sales) that may not be explained by the trend or seasonal components.

In addition to helping to understand and interpret economic data, seasonal decomposition can also help forecast future economic trends. By analyzing the trend and seasonal components of the data, economists can make informed predictions about how the economy is likely to perform in the future.

Figure 2.2 shows the additive seasonal decomposition of the USA's Consumer Price Index data. Trend data can be interpreted as a positive trend constantly growing over time. Seasonal trend states the rise and falls of the data in given periods; this is important because below, we can see that the economy –throughout time– tends to follow the same increase and decrease in the given periods every 12 months. The residual can be considered as the market's error or volatility. It can be seen that the market was highly volatile during the USA's recession of 2008. This is not necessary, but it is always good practice to understand the data in better detail.

Overall, additive seasonal decomposition is a valuable tool for understanding and interpreting time series data and identifying patterns and trends that may take time to be apparent in the raw data.

Figure 2.2

AUGMENTED DICKEY-FULLER TEST (ADF)

To perform the ADF test, economists need to specify a lag order, which determines the number of lags to include. The test statistic is then calculated based on the data and the specified lag order, and a p-value is generated based on the test statistic. If the p-value is below a specified threshold (typically 0.05), it indicates that the null hypothesis can be rejected and that the data is stationary.

To be stationary, time series data must have a constant mean, variance, and autocorrelation structure over time. If time series data is non-stationary, it may exhibit trend or seasonal patterns, making it challenging to model accurately and forecast.

Economists often use the ADF test to determine whether economic data, such as GDP, inflation, employment, or trade, is stationary or non-stationary. If the data is non-stationary, economists may need to transform it (such as differencing or logging) to make it stationary before applying a modeling technique like ARIMA.

The ADF test is a hypothesis test that is used to determine whether a time series data is stationary or non-stationary. The test works by testing the null hypothesis that the data is non-stationary against the alternative hypothesis that it is stationary. If the null hypothesis is rejected, it suggests that the data is stationary and can be used for modeling and forecasting.

To perform the ADF test, economists need to specify a lag order, which determines the number of lags to include in the test. The test statistic is then calculated based on the data and the specified lag order, and a p-value is generated based on the test statistic. If the p-value is below a specified threshold (typically 0.05), it indicates that the null hypothesis can be rejected and that the data is stationary.

Overall, the ADF test is a widely used and accepted method for determining the stationarity of time series data. It is beneficial for economists to analyze and forecast economic indicators. By understanding whether time series data is stationary or non-stationary, economists can choose the appropriate modeling techniques and make more accurate forecasts.

In this data, with an ADF statistic value of 3.3 and a p-value of 1, the null hypothesis of non-stationarity cannot be rejected. This means there is insufficient evidence to conclude that the time series is stationary. The time series may be non-stationary, but more data or a different statistical test may be needed to confirm this.

KWIATKOWSKI-PHILLIPS-SCHMIDT-SHIN (KPSS) TEST

The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is a statistical test that is used to determine whether a time series data is stationary. Stationarity is an important assumption for many time series modeling techniques, including the autoregressive integrated moving average (ARIMA) model.

To be stationary, time series data must have a constant mean, variance, and autocorrelation structure over time. If time series data is non-stationary, it may exhibit trend or seasonal patterns that can make it challenging to model and forecast accurately.

Economists often use the KPSS test to determine whether economic data, such as GDP, inflation, employment, or trade, is stationary or non-stationary. If the data is non-stationary, economists may need to transform it (such as differencing or logging) to make it stationary before applying a modeling technique like ARIMA.

The KPSS test is a hypothesis test used to determine whether a time series data is stationary or non-stationary. The test works by testing the null hypothesis that the data is stationary against the alternative hypothesis that it is non-stationary. If the null hypothesis is rejected, it suggests that the data is non-stationary and may need to be transformed (such as through differencing or logging) to make it stationary before applying a modeling technique like ARIMA.

To perform the KPSS test, economists need to specify a lag order for the test, which determines the number of lags to include in the test. The test statistic is then calculated based on the data and the specified lag order, and a p-value is generated based on the test statistic. If the p-value is below a specified threshold (typically 0.05), it indicates that the null hypothesis can be rejected and that the data is non-stationary.

The KPSS test is a widely used and widely accepted method for determining the stationarity of time series data, and it is beneficial for economists who are analyzing and forecasting economic indicators. By understanding whether time series data is stationary or non-stationary, economists can choose the appropriate modeling techniques and make more accurate forecasts.

In this data, with a KPSS statistic value of 4.96508 and a p-value of 0.01, the null hypothesis of stationarity cannot be rejected. This means there is insufficient evidence to conclude that the time series is non-stationary. It is possible that the time series is stationary, but more data or a different statistical test may be needed to confirm this.

BOX-COX TRANSFORMATION

The Box-Cox transformation is a statistical technique used to transform non-normal data into a routine or nearly normal distribution. It is often used in conjunction with time series modeling techniques, such as the autoregressive integrated moving average (ARIMA) model, to stabilize the data's variance and make it more suitable for modeling and forecasting.

The Box-Cox transformation can be used to stabilize the variance of the data by transforming the data into a routine or nearly normal distribution. This can be particularly useful for data that exhibits heteroscedasticity or non-constant variance.

To perform a Box-Cox transformation, you need to specify a lambda value, which determines the degree of the transformation. A lambda value of 0 corresponds to a log transformation, while a lambda value of 1 corresponds to no transformation. The transformation is applied to the data using the following formula:

y = (x^lambda - 1) / lambda

Where y is the transformed data, x is the original data, and lambda is the specified transformation parameter.

The appropriate lambda value for a given dataset can be determined using statistical tests and techniques, such as the Anderson-Darling test or the Kolmogorov-Smirnov test. Once the appropriate lambda value has been determined, the Box-Cox transformation can be applied to the data to stabilize the variance and make it more suitable for modeling and forecasting.

Overall, the Box-Cox transformation is a valuable tool for transforming non-normal data into a routine or nearly normal distribution. It is often used in conjunction with time series modeling techniques like ARIMA to improve the accuracy of forecasts. The results can be seen below. An interesting behavior of the box cox is that after the transformation, the values are almost the same as inflation.

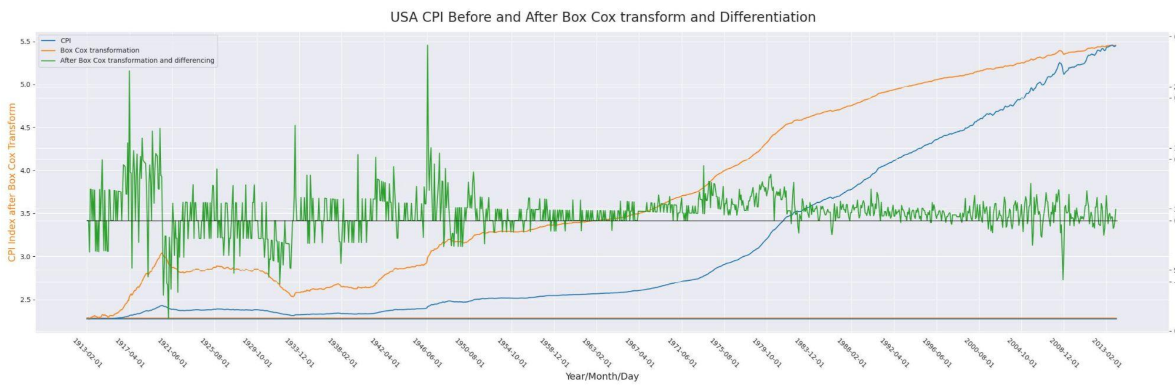


Figure 2.3

AUTOCORRELATION

Autocorrelation is a statistical measure that describes the correlation between the values of a time series with itself at different time lags. It is often used to identify the degree to which a time series is dependent on its past values.

To prepare data for an ARIMA model using autocorrelation, you can use statistical tools such as the autocorrelation function (ACF) and partial autocorrelation function (PACF) to visualize the degree of autocorrelation in the data at different lag periods. The ACF plots the autocorrelation of the data at different lag periods. In contrast, the PACF plots the partial autocorrelation of the data, which is the autocorrelation of the data with the effects of the intermediate lags removed.

The ACF and PACF plots can help you identify patterns and trends in the data, such as seasonality or autoregressive patterns, and can inform the choice of p and q values for the ARIMA model. For example, suppose the ACF plot shows a significant positive autocorrelation at lag 1. In that case, it may suggest that an AR(1) model is appropriate, while if the PACF plot shows a significant positive autocorrelation at lag 2, it may suggest that a MA(2) model is appropriate.

Once the appropriate values for the p, d, and q parameters have been determined using the AC, you can then proceed with fitting an ARIMA model to the transformed data. This can be useful for forecasting future values of the time series or for analyzing trends in the data. The values obtained for the ARIMA model are p: 6, d:1, and q:15.

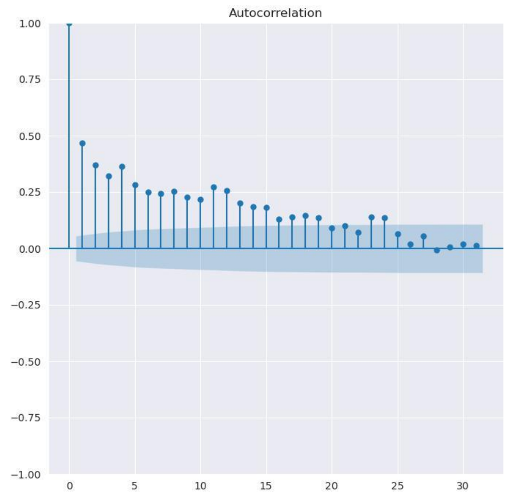


Figure 2.4

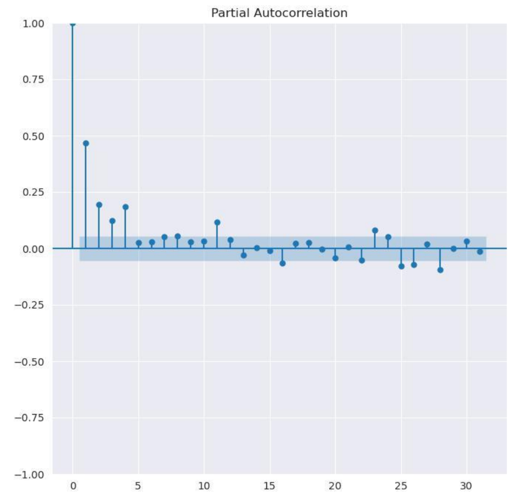


Figure 2.5

AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA)

The autoregressive integrated moving average (ARIMA) model is a statistical model used to analyze and forecast time series data. It is a widely used economic model for analyzing and forecasting economic indicators such as GDP, inflation, employment, and trade.

An ARIMA model consists of three components: autoregressive (AR), integrated (I), and moving average (MA). The AR component models the dependence of the current value of the time series on past values, the I component models the effect of differencing the data to make it stationary, and the MA component models the effect of past errors or shocks on the current value of the time series.

To use an ARIMA model, economists first need to determine the appropriate values for the p (AR), d (I), and q (MA) parameters based on the characteristics of the data. This can be done using statistical tests and techniques such as the Augmented Dickey-Fuller (ADF) test, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, and the autocorrelation function (ACF) and partial autocorrelation function (PACF).

Once the appropriate values for the p, d, and q parameters have been determined, the economist can fit the ARIMA model to the data and use it to make forecasts. The model can be used to forecast a single point in the future or to generate a forecast for several future periods.

Economists often use ARIMA models to forecast economic indicators and understand the underlying factors driving the data. For example, an economist might use an ARIMA model to forecast a country's GDP growth and identify the factors contributing to the economy's growth or contraction.

Overall, the ARIMA model is a powerful and widely used tool in economics for analyzing and forecasting time series data, particularly economic indicators. It is particularly useful for understanding and predicting trends and patterns in data and identifying the underlying factors driving the data.

Figure 2.6 is the culmination of all the work done beforehand. After all this, it can be seen that the predictions of almost two years are the main setback of the ARIMA model. The problem is that the ARIMA model is nearsighted and cannot look too far into the future due to the very little data used.

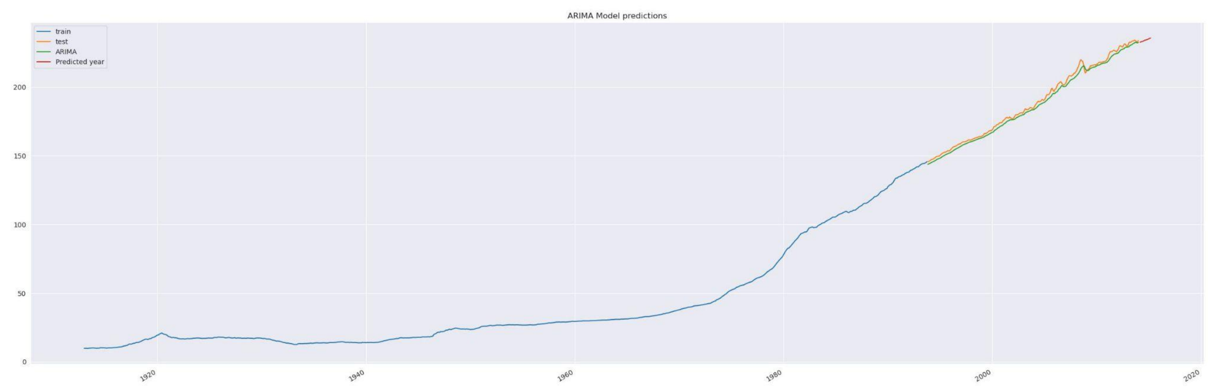


Figure 2.6

CONCLUSION

In conclusion, the Consumer Price Index (CPI) is a measure used to track changes in the cost of living and adjust for inflation. *Inflation* is the rate at which the general level of prices for goods and services is rising. The main goal of predicting the CPI and inflation is to anticipate changes in the cost of living and provide policymakers and businesses with a sense of the direction in which prices are moving. One tool commonly used to prepare data for analysis is a box plot, and ARIMA is a statistical method that can be used to analyze and forecast time series data. However, it is essential to note that the results of these predictions may only sometimes be accurate, as many factors can influence the movements of the CPI and inflation. These may include economic conditions, supply and demand changes, and economic and fiscal policy shifts. As a result, it is essential to consider all relevant information when making predictions carefully and to recognize that the results may not always be reliable.

BIBLIOGRAPHY

DanlBradley. (n.d.). *Danlbradley/recessionpredictor: Predicts whether or not there will be a recession in the next fiscal quarter*. GitHub. Retrieved December 31, 2022, from https://github.com/DanlBradley/RecessionPredictor

Bradley, D. (2022, April 28). *Forecasting recessions with Scikit-Learn*. Medium. Retrieved December 31, 2022, from https://medium.com/mlearning-ai/forecasting-recessions-with-scikit-learn-df58e1ea695f

USA CPI Dataset: https://datahub.io/core/cpi-us

Berg, A., & Pattillo, C. (2000, July). *Economic issues no. 22 -- the challenge of predicting economic crises*. International Monetary Fund. Retrieved November 11, 2022, from https://www.imf.org/external/pubs/ft/issues/issues22/

Zhang, T. (2021, January 11). *Recession prediction using machine learning*. Medium. Retrieved December 31, 2022, from https://towardsdatascience.com/recession-prediction-using-machine-learning-de6eee16ca94