INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Predicting Economic Recessions Using Machine Learning an Analysis of Macroeconomic Indicators

JOÃO NUNO BORGES PINTO RAMOS
Setembro de 2025

POLITÉCNICO
DO PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

Predicting Economic Recessions Using Machine Learning

an Analysis of Macroeconomic Indicators

João Nuno Borges Pinto Ramos

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Science, Specialisation Area of Data Engineering

Supervisor: Maria de Fátima Coutinho Rodrigues (mfc@isep.ipp.pt)

Porto, September 2025

Acknowledgments

This section is used to express my profound gratitude to all the people that were involved and played a significant role in the development of this dissertation.

First, I want to thank my supervisor, Professor Maria de Fátima Rodrigues (MFC), for the continuous support and willingness to help at any moment during the development of the project. Her guidance, constructive feedback, and readiness to provide suggestions, even in the final stages, were invaluable in improving this dissertation.

Second, I want to thank my family for their unwavering support, patience, and encouragement throughout this journey. A very special thanks goes to my mother, my brother, and to those who, unfortunately, are no longer part of my life but, I am certain, would be proud of me in this achievement.

Thirdly, a word of gratitude to my work colleagues, for their understanding and support throughout this period. Balancing the demands of work and the writing of this dissertation was not easy, and I deeply appreciate having colleagues whose presence and encouragement helped ease the challenge.

This dissertation serves as a testimony of resilience, hard work, and persistence, and at this moment, for fear of overlooking someone, I wish to extend my gratitude to all who, in one way or another, contributed to the development of this project.

Statement of Integrity

I hereby declare that I have conducted this academic work with integrity.

I have not plagiarised or applied any form of undue use of information or falsification of results along the process leading to its elaboration.

Therefore, the work presented in this document is original and authored by me, having not previously been used for any other end. The exceptions are explicitly recognised in the section “Ethical considerations” of the first chapter. This section also states how AI tools were used and for what purpose.

I further declare that I have fully acknowledged the Code of Ethical Conduct of P.PORTO.

ISEP, Porto, 28 de September de 2025

João Nuno Borges Pinto Ramos

Resumo

As recessões económicas têm implicações profundas na coesão e bem-estar social. Apesar da sua frequência e duração terem diminuído nas últimas décadas, continuam a ser um fenómeno recorrente tanto em economias avançadas como emergentes.

Os modelos econométricos tradicionais, embora relevantes, revelam limitações na capacidade de captar as interações complexas e não lineares entre variáveis macroeconómicas, que estão na origem de recessões económicas.

O trabalho desenvolvido nesta dissertação tem como principal foco a análise e comparação do desempenho de diferentes abordagens de aprendizagem automática no contexto da previsão da ocorrência de recessões económicas, em particular, no contexto da economia dos Estados Unidos da América.

Para este efeito, vários indicadores macroeconómicos, referentes a várias categorias da economia americana, foram recolhidos do repositório de dados da Reserva Federal de St.Louis (FRED), desde 1959 até 2024. Após uma análise estatística detalhada e a definição de diferentes estratégias de seleção de variáveis, como por exemplo, Correlação de Pearson e Análise de Componentes Principais (PCA), foram implementados e avaliados modelos estatísticos e de aprendizagem automática em três diferentes períodos específicos de recessão da economia americana: 1973-1976, 1980-1983 e 2007-2010.

Os resultados obtidos demonstram que os modelos SARIMAX e Regressão Linear apresentaram um melhor desempenho em relação aos restantes modelos. Para além disto, verificaram-se melhorias significativas no desempenho dos modelos, quando se aplicam estratégias de seleção de variáveis, com especial enfoque na estratégia de Correlação de Pearson.

Em resumo, a presente dissertação contribui para o avanço na investigação na área de previsão de recessões económicas, através de uma comparação metodológica entre diversos tipos de modelos estatísticos e de aprendizagem automática, no contexto específico de três períodos históricos de recessão na economia americana.

Palavras-chave: Previsão, Recessão Económica, Aprendizagem Automática, Inteligência Artificial

Abstract

Economic recessions have profound implications on the social cohesion and overall well-being of society. Despite their declining frequency and duration in recent decades, recessions remain a recurring phenomenon in both advanced and emerging economies. Traditional econometric approaches, while valuable, often struggle to account for the multifaceted and nonlinear interactions between macroeconomic variables that drive recessionary dynamics. This dissertation examines how different machine learning models behave in the context of forecasting the occurrence of economic recessions in the U.S. economy. For this purpose, several macroeconomic series from the U.S. economy were collected from the FRED-QD database. Feature selection strategies, such as Pearson correlation and Principal Component analysis, were applied to address issues of high dimensionality, and a range of statistical and machine learning model, including SARIMAX, Linear Regression, Random Forest and XGBoost were implemented and tested across three historical recession periods (1973-1976, 1980-1983 and 2007-2010). The results show that SARIMAX and Linear Regression consistently outperform the other models across the three different testing periods. Furthermore, the findings emphasize the importance of feature selection strategy, with Pearson Correlation filtering enabling improved predictive performance when compared with other strategies.

Keywords: Prediction, Economic Recession, Machine Learning, AI

Table of Contents

List of Figures

List of Tables

Abbreviations and Symbols

AI	**Artificial Intelligence**
BMA	**Bayesian Moving Average**
CPI	**Consumer Price Index**
CNN	**Convolutional Neural Network**
FCNN	**Fully Connected Neural Networks**
FFNN	**Feed Forward Neural Networks**
FRED	**Federal Reserve Economic Data**
GDP	**Gross Domestic Product**
GRU	**Gated Recurrent Unit**
IPP	**Instituto Politécnico do Porto**
KNN	**K-Nearest Neighbours**
LSTM	**Long Short-Term Memory**
ML	**Machine Learning**
MLP ANN	**Multilayer Perceptron Artificial Neural Networks**
OLS	**Ordinary Least Squares Regression**
RF	**Random Forest**
USD	**United States Dollar**
LWTA	**Local Winner-Takes-All**

1 Introduction

This introductory chapter aims to present the research context, define the problem under study, and outline the objectives and methodology adopted. It concludes by addressing the ethical and legal considerations observed throughout the project and by briefly introducing the overall structure of the dissertation.

1.1 Context

The National Bureau of Economic Research (NBER) defines an economic recession as a "significant decline in economic activity that is spread across the economy and lasts more than a few months" (National Bureau of Economic Research, 2024). Recessions have profound impacts on economies, often resulting in increased unemployment rates and inflation. The deterioration of economic well-being extends beyond financial aspects, affecting areas such as social cohesion, criminal activity (Brian Bell, 2015) or mental health (Guerra and Eboreime, 2021).

Although the frequency of recessions has gone down over the last few decades, economic downturns still happen with a considerable frequency. According to the International Monetary Fund, between 1960 and 2007, 21 advanced economies were in recession 10% of the time (Claessens and KOSE, n.d.). In the US, from 1857 to 2020, there have been 34 recessions, occurring on average every 4.8 years and lasting 17 months. Notably, economic recessions in the US have become shorter and less frequent in the last 40 years. From 1980 to 2020, there have been 6 recessions, occurring on average every 6.7 years and lasting 10 months (NBER, 2023). In Portugal, the economic cycles follow a similar pattern, between 1980-2022, the Portuguese economy has experienced 6 recessions (Aguiar-Conraria *et al.*, 2023).

Given the frequency and the social impacts of economic recessions, it is important to study the underlying causes and the influence that economic indicators can have on their prediction. Machine Learning models, with their ability to capture complex patterns in data, can offer a promising alternative for predicting the occurrence of recessions.

1.2 Problem

The prediction of recessions is particularly challenging due to the multifaceted nature of their causes, that can include a combination of supply and demand shocks, market dynamics, policy changes, financial crises or housing market crashes (Weinstock, 2023). These complex interactions make it difficult to identify a single set of indicators or models that can reliably predict economic downturns.

Traditional economic models, while valuable, may fail to capture complex patterns. ML models can identify hidden patterns and adapting to changes in economic conditions. This ability positions ML as a promising tool for improving the accuracy and timeliness of recession forecasts.

1.3 Research Objectives

The main objective of this project is to develop a Machine Learning model capable of effectively predict the onset of an economic recession. To achieve this, a series of specific, actionable objectives will guide the development process:

- Collect and preprocess macroeconomic data from reliable sources over several decades.
- Identify the most critical indicators for predicting economic recessions using feature selection techniques.
- Develop and compare different machine learning models for predicting the onset of economic recessions.
- Evaluate the models' performance and select the best-performing model.
- Provide insights into how different economic factors contribute to the risk of a recession, aiding policymakers and investors in decision-making.

1.4 Methodology

In order to achieve the research objectives laid out previously, a data science project will be developed to assess the effectiveness of machine learning models in recession forecasting. The main methodology framework that guided the development of this project is the CRISP-DM methodology. Initially proposed in 1996, but described in detail in (Chapman *et al.*, 2000), this methodology encompasses six iterative and interconnected phases.

Within the scope of this dissertation, the first five phases were fully addressed, while the final phase (Deployment) was not included, as the focus remained on methodological development and evaluation instead of deploying the selected model in a production environment:

- **Business Understanding**: Defined the main objective of the study, namely, to explore how machine learning methods can be applied to forecast economic recessions, translating this into measurable research objectives. This involved

studying in detail the state-of-the-art literature to assess how forecasting is performed in an economic context.

- **Data Understanding**: Collected the dataset from FRED-QD database, described its structure and content, and performed exploratory data analysis to identify patterns, assess data quality and detect issues.
- **Data Preparation**: Applied relevant transformation and cleaning steps to the raw dataset. These included the removal of variables with missing values and the standardization of the explanatory variables. After that two different feature selection strategies were implemented in the raw dataset, Pearson Correlation and Principal Component Analysis (PCA).
- **Modelling**: In this phase, Linear Regression, SARIMAX, Random Forest and XGBoost were selected as the forecasting models to be assessed. The forecasting strategy was defined and applied, included data partitioning methodology, splitting the dataset into training, validation and testing sections. A rolling origin cross-validation was then carried out to tune different hyperparameters in the different models and to check their performance on the validation data
- **Evaluation**: Finally, the best model configurations, selected through the validation data evaluation strategy, were tested on unseen data across three different testing periods. Model performance was assessed using two regression metrics, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE), in order to determine which approach was the most effective across the different testing windows.

1.5 Ethical and Legal Considerations

The IPP's Code of Conduct ("CÓDIGO DE CONDUTA — P.PORTO | Ensino Superior Público", n.d.) was rigorously adhered to throughout the development of this document. This Code provides valuable insights into the ethical duties and responsibilities of students and the broader academic community, explicitly condemning any form of academic misconduct as outlined in Article 6, Line n). This includes plagiarism, unauthorized collaboration, and fraudulent practices.

Throughout the development of this document, there was a consistent effort to comply with the Code of Conduct. For instance, the state-of-the-art review was conducted using the PRISMA guidelines, ensuring a thorough and transparent review process. All references were cited to uphold academic integrity and appropriately credit the original authors.

Additionally, the European Commission's Ethics Guidelines for Trustworthy AI ("Ethics guidelines for trustworthy AI | Shaping Europe's digital future", n.d.) were followed to ensure the AI system is developed in a manner that is ethical, transparent, and aligned with human-centred values.

When it comes to permissions and approvals, the project was developed using publicly available economic data. The technical development of the project was used exclusively with open-source libraries and programming languages. As such, there wasn't a need to ask for approval for specific data access.

Finally, regarding the use of Artificial Intelligence tools, it is important to acknowledge that a large language model (LLM) was employed in some specific parts during the project. Its usage was limited to the following purposes:

- **Reviewing and organizing text**: The LLM was used exclusively for editing, improving clarity, and organizing content, without contributing to the generation of ideas or conceptual development.
- **Supporting the development phase**: The LLM assisted in the coding process, primarily by helping to resolve bugs and generate boilerplate code.
- **Technical formatting and presentation**: The LLM was also used to streamline and refine the preparation of tables, plots, figures and captions.

Overall, the use of the LLM was restricted to auxiliary tasks, guiding the process and enhancing productivity and presentation quality. All methodological decisions, analytical procedures and interpretations were entirely conceived, implemented and validated by the author of the project.

1.6 Document Structure

This dissertation is structured to align with the CRISP-DM methodology, with each chapter aiming at representing phases of the methodological framework. The document is therefore organized as follows:

- **Chapter 2 – Background:** The chapter aims at introducing core concepts of time series forecasting, covering forecasting strategies and an overview of statistical and machine learning models.
- **Chapter 3 – State-of-the-Art Review**: Provides an overview of existing literature on recession forecasting, addressing the key techniques and economic indicators used in machine learning approaches to predicting economic recessions.
- **Chapter 4 – Exploratory Data Analysis**: This chapter presents a detailed analysis of the dataset, focusing on patterns, correlations and other insights relevant to modelling.
- **Chapter 5** – **Data Preparation and Modelling**: This chapter describes the steps required to prepare and preprocess the dataset, followed by the application of feature selection strategies. It then introduces the forecasting models employed and describes the overall forecasting methodology adopted.
- **Chapter 6** – **Results and Discussion**: This chapter aims at presenting and comparing the forecasting results, highlighting key findings, and reflecting on limitations and considerations.
- **Chapter 7** – **Conclusion**: This chapter summarizes the findings in the scope of the research objectives, while suggesting directions for future research.

2 Background

This chapter introduces theoretical background knowledge relevant to this study. It begins by framing the machine learning task of recession forecasting, followed by an overview of key time series concepts and forecasting strategies. The chapter then continues, presenting the main modelling approaches, covering both statistical and machine learning models. Finally, some feature selection strategies are introduced.

2.1 Framing the Machine Learning Task: Forecasting GDP

As described in the first chapter of this thesis, namely in the subchapter 1.3, the main objective of this study is to explore how machine learning methods can be applied to effectively forecast economic recessions using macroeconomic indicators. This machine learning task can be framed in one of two ways: as a binary classification problem, where the model predicts whether a future time period will be a recession or not; or as a regression problem, where the main objective is to predict a numeric value of a economic indicator, usually the GDP, and then infer the recession state from this behaviour.

It is, therefore, important to mention that, in this study, the problem is framed as a time series regression task, with the model developed with the goal of predicting the quarterly real GDP based on a range of macroeconomic indicators.

The decision to frame the problem as a regression task instead of a binary classification problem is motivated by several practical considerations. First, this approach simplifies implementation and aligns with the continuous nature of macroeconomic indicators. Second, regression provides a more informative supervision signal than classification, allowing the model to learn from both the magnitude and direction of GDP changes over time, which is particularly useful when working with limited observations of recession periods. Moreover, evaluating regression models using standard metrics such as RMSE or MAE is straightforward and avoids the complexity of tuning classification thresholds. Third, treating the problem as a regression task enhances generalizability across regions or datasets, especially in cases where recession labels may be unavailable, inconsistent, or defined differently. Finally, real GDP figures are consistently published and regularly updated, making them a reliable and objective target variable for training predictive models across various time periods and countries.

2.2 Time Series Overview

Given that this thesis frames the problem as a timeseries machine learning task, it is essential to first have a clear understanding of the fundamental concepts of time series, including their definition, key characteristics and common forecasting strategies. The knowledge presented in this subchapter will support the reader in understanding the technical decisions described in the subsequent chapters.

2.2.1 Characteristics of Time Series

Time series can be defined as sequentially observed data points over time (Chatfield and Xing, 2019). As described by Cristopher Chatfield, the statistical analysis of time series can be divided into six main categories: Economic and Financial Time Series, Physical Time Series, Process Control Data, Binary Processes, Marketing Time Series and Point Processes.

Timeseries data can be decomposed into four key components:

- **Trend**: Represents the long-term direction of a time series in the absence of any other variation.
- **Seasonality**: Variation in the time series that occurs at fixed time periods over time, such as every year in the same months or every week on the same days.
- **Cyclicality**: Represent large variations in the data that recur over longer time periods than seasonal variations, but not with a fixed frequency.

- **Noise/Residuals**: Represents random, unpredictable variations that cannot be explained by any other component listed, such as trend or seasonality. Can also be described as white noise.

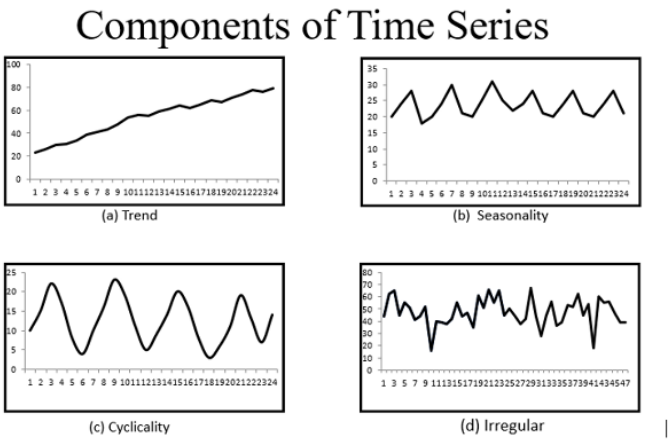

Figure 1 - Components of Time Series. Source: (Shailesh, 2024)

As described by (Mathelinea *et al.*, 2023), one of the most important steps in time series analysis and forecasting, particularly when using statistical models, is to convert nonstationary data into stationary data using techniques such as differencing. It is therefore essential to understand the concepts of stationarity and differencing.

Stationarity is the condition in which a time series exhibits, over different time periods, constant mean, constant variance or standard deviation and that the autocorrelation does not change over time. In other words, in mathematical terms, a time series is stationary when its statistical properties are independent of time (Hyndman and Athanasopoulos, 2021).

One of the most common ways of assessing the stationarity of a specific time series is to perform a unit root test. There are two widely used statistical hypothesis tests for this purpose: the **Augmented Dickey-Fuller (ADF)** test and the **Kwiatkowski-Phillips-Schmidt-Shin (KPSS)** test. In the Augmented Dickey-Fuller test, the null hypothesis states that the time series is non-stationary. Rejecting the null hypothesis proves that the time series is stationary (Dickey and Fuller, 1979). In contrast, the KPSS test takes a different approach: its null hypothesis assumes that the series is stationary around a deterministic trend (Kwiatkowski *et al.*, 1992). Therefore, rejecting the KPSS null suggests non-stationarity.

As mentioned, differencing is the process of converting nonstationary time series into stationary data. According to (Hyndman and Athanasopoulos, 2021), differencing can be achieved by computing the differences between consecutive observations in the original series. It can be described in the following function:

$$y_t' = y_t - y_{t-1}$$

In the cases where the time series remains nonstationary after **first-order differencing**, the process can be repeated on the already differenced series, therefore applying a **second-order differencing**.

2.2.2 Forecasting Concepts and Strategies

2.2.2.1 One-step vs Multi-step Forecasting

In **one-step forecasting**, the model is trained to predict the next immediate point based on past observations.

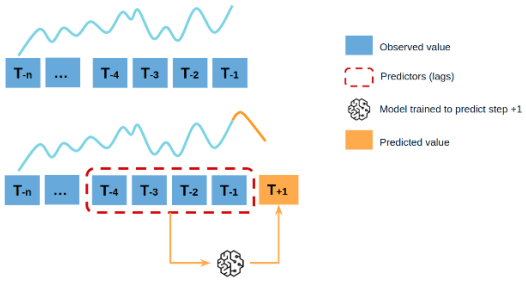

Figure 2 - Diagram of single-step forecasting. Source: (Rodrigo and Ortiz, 2024a)

Conversely, **multi-step forecasting** refers to predicting multiple time steps into the future. The number of steps ahead is commonly referred to as the **forecasting horizon**. In the context of the problem addressed in this document, the forecasting horizon may correspond, for example, to one, two, or four quarters ahead, depending on the intended application and planning requirements.

2.2.2.2 Recursive vs Direct Forecasting

Multi-step forecasting can be achieved via two main strategies: **Recursive** and **Direct** Forecasting.

In **Recursive Forecasting**, a single model is trained to perform one-step forecasts, and its own predictions are reused, recursively, as inputs, to generate further steps. The process continues until all the future steps are produced (Rodrigo and Ortiz, 2024a). Prediction errors can propagate and amplify with each step.

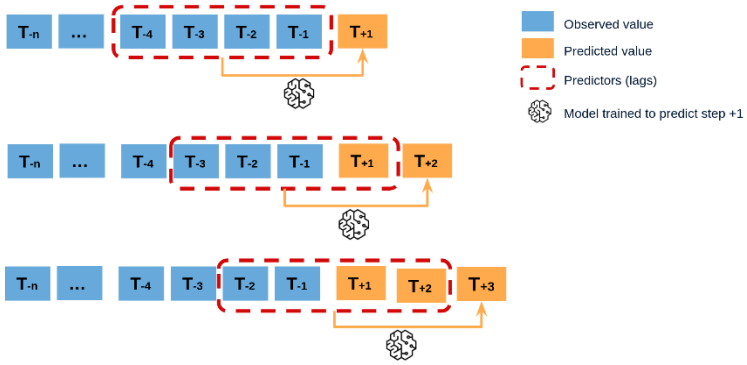

Figure 3 - Recursive Multi-step forecasting. Source:(Rodrigo and Ortiz, 2024a)

In contrast, **Direct Forecasting** consists of training a different independent model for each step of the forecast horizon. For example, if the objective is to predict the next five values of a given time series, five different models are trained, one for each step. As a result, each forecast is made independently, without relying on the predictions of the previous steps (Rodrigo and Ortiz, 2024a). Often more accurate for longer horizons, but more computationally expensive.

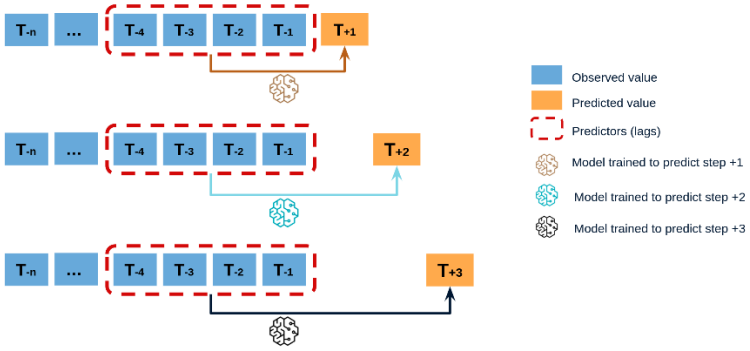

Figure 4 - Direct Multi-step forecasting. Source:(Rodrigo and Ortiz, 2024a)

In addition to Recursive and Direct forecasting, several hybrid strategies, such as DirRec, MIMO, and DIRMO, have been proposed for multi-step time series forecasting. These strategies combine aspects of both approaches to balance accuracy, efficiency, and error control.

2.2.2.3 Expanding vs Rolling/Sliding Window Evaluation

One of the most common methodologies to evaluate the predicting capabilities of a given model on a time series over time, is the use of expanding or sliding/rolling window evaluation techniques. These approaches simulate real-world forecasting scenarios by training the model on sequential subsets of data and testing it on future observations.

Expanding Windows have a fixed starting point and incorporate new data as it becomes available (Luka, 2020). In this approach, the training window expands over the entire history of a time series and is tested repeatedly against the forecasting window without dropping older

data points (Bell and Smyl, 2018). This technique is particularly useful when dealing with small datasets/short series, as it makes use of the historical information for model training.

Figure 5 - Expanding Window. Source: (Bell and Smyl, 2018)

In contrast, in **Sliding Window** involves a fixed-size training window that moves forward through the time series as new data becomes available. At each of the steps, the model is trained on a subset of the time series with a fixed length. As the window advances, older data points are discarded, and new ones are included.

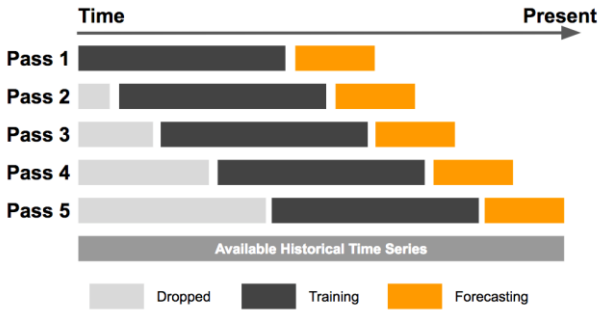

Figure 6 - Sliding Window. Source: (Bell and Smyl, 2018)

2.3 Feature Selection Strategies

Feature selection strategies play an important role in optimizing and improving the performance of machine learning models. Their main purpose is to reduce the dimensionality of the dataset, mitigating issues such as overfitting, high computational cost, and noise from irrelevant variables. In this subchapter, two feature selection strategies (Pearson Correlation and PCA) applied in the scope of this project are presented, with the objective of providing the necessary background for their use.

2.3.1 Pearson Correlation

A common feature selection strategy for dimensionality reduction is the use of the Pearson Correlation method, which calculates the correlation coefficient between two numerical variables. In the context of a machine learning task, this involves computing the correlation between each candidate feature and the numerical target variable and retaining the most correlated features based on a predefined threshold. The Pearson Correlation Coefficient measures both the strength and direction of a linear relationship, with values ranging from -1 to +1 (Kenton et al., 2024):

- +1 → perfect positive linear association (as GDP increases, the variable increases proportionally).
- -1 → perfect negative linear association (as GDP increases, the variable decreases proportionally).
- 0 → no linear association.

2.3.2 Principal Component Analysis (PCA)

Developed in 1901 by Karl Pearson, Principal Component Analysis (PCA) is a technique that reduces the number of dimensions in large datasets to principal components that retain most of the original information. It works by transforming potentially correlated variables into a smaller set of variables, called principal components ("What Is Principal Component Analysis (PCA)? | IBM", n.d.).

The PCA process can be summarized in a sequence of computational steps ("What Is Principal Component Analysis (PCA)? | IBM", n.d.):

- **Standardize the data**: center each variable by subtracting its mean and scale by its standard deviation so all variables are on the same scale.
- **Compute the covariance matrix**: summarize the relationships between variables to see how they vary together.
- **Find eigenvalues and eigenvectors** – Eigenvectors (principal components) show directions of maximum variance, and eigenvalues indicate how much variance each component explains.

As ("What Is Principal Component Analysis (PCA)? | IBM", n.d.) describes, the first principal component (PC1) calculated refers to the direction in space along which the data points have the highest or most variance. It is, therefore, the line that best represents the shape of the projected points. The larger the variability captured in the first component, the larger the information retained from the original dataset

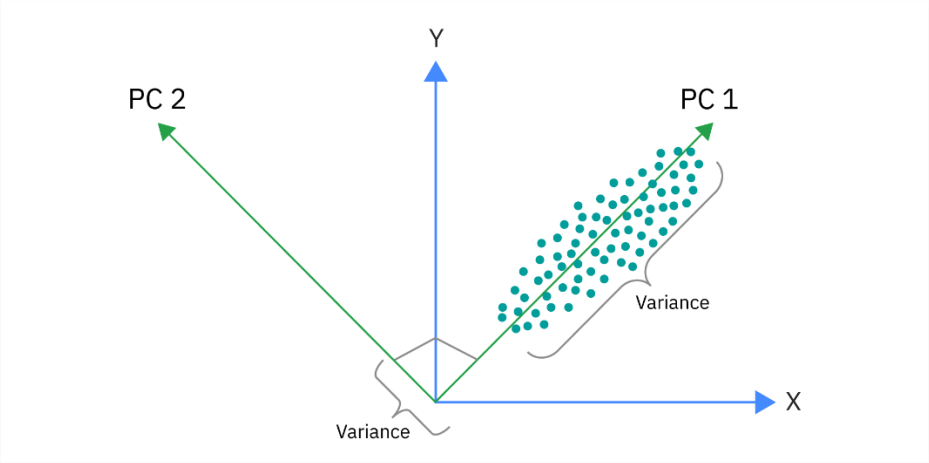

Figure 7 - Overview of PCA Components. Source: ("What Is Principal Component Analysis (PCA)? | IBM", n.d.)

2.4 Overview of Modelling Approaches

This subchapter presents an overview of the main modelling approaches used in time series forecasting, grouped into two categories: **Statistical Models** and **Machine Learning Models**.

2.4.1 Statistical Models

2.4.1.1 ARIMA Models

ARIMA stands for Auto Regressive Integrated Moving Average. It is the combination of three different components: **Differencing** (converting non-stationarity time series into a stationary time series), **Autoregression** models and **Moving Average** models.

In a Autoregression Model (AR), the target variable is forecasted using a linear combination of past values of the variable. The term of "autoregression" implies that the target variable is regressed on its own prior values. An AR model of order p can be expressed as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

In contrast to AR models, a Moving Average model (MA) uses past forecast errors in a regression-like model. An MA model with order q can be defined as:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

Because both AR and MA models assume that the time series is stationary to make predictions, it is necessary to apply differencing of order d. Therefore, as initially mentioned, the ARIMA (p,d,q) model is the combination of three elements:

- **p**: the order of the autoregressive component (lag order).
- **d:** degree of differencing.
- **q:** order of the moving average component.

ARIMA models are also capable of modelling timeseries with seasonal characteristics. Seasonal ARIMA (SARIMA) integrates an additional m seasonal component in the original ARIMA model previously described, to indicate the number of observations per year. SARIMA model can be defined as follows:

$$\text{ARIMA}\quad \underbrace{(p,d,q)}_{\substack{\uparrow\\ \text{Non-seasonal part}\\ \text{of the model}}} \quad \underbrace{(P,D,Q)_m}_{\substack{\uparrow\\ \text{Seasonal part of}\\ \text{of the model}}}$$

Furthermore, the ARIMA model can also be extended to account for both **seasonal patterns** and **external explanatory variables**, resulting in the **SARIMAX** model (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors). The inclusion of external feature parameters can improve the model's performance by reducing the prediction errors and addressing the autocorrelation issues in residuals (Alharbi and Csala, 2022) .

2.4.1.2 Vector Autoregression (VAR) Model – Multivariate Forecasting

A **Vector Autoregression (VAR)** is a multivariate time series model that captures the linear interdependencies between multiple time series. It is considered a generalization of AR(p) model for forecasting a vector of time series (Hyndman and Athanasopoulos, 2021).

A two variable VAR(1) model with one lag can be described with the following equation:

$$\begin{aligned} y_{1,t} &= c_1 + \phi_{11,1} y_{1,t-1} + \phi_{12,1} y_{2,t-1} + e_{1,t} \\ y_{2,t} &= c_2 + \phi_{21,1} y_{1,t-1} + \phi_{22,1} y_{2,t-1} + e_{2,t}, \end{aligned}$$

In this equation, $e_{1,t}$ and $e_{2,t}$ represent the white noise processes that may be correlated. Coefficient $\phi_{ii,l}$ captures the influence of the ℓth lag of variable y_i on itself, while the coefficient $\phi_{ij,l}$ captures the influence of the ℓth lag of variable y_j on y_i (Hyndman and Athanasopoulos, 2021)

When it comes to stationarity, it is important to note that, like univariate models, such as ARIMA, VAR models require the variables to be stationary time series.

2.4.2 Machine Learning Models

2.4.2.1 Linear Regression

A simple Linear Regression models the linear relationship between the forecast variable y and a single predictor variable x. In the context of time series, linear regression models are used to forecast the time series of interest y, assuming a linear regression with other time series x (Hyndman and Athanasopoulos, 2021). The following equation describes a single predictor linear regression:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t.$$

If there are two or more predictors, the model is designated as **multiple regression model**. The following equation describes a multiple regression model:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t,$$

According to (Hyndman and Athanasopoulos, 2021), Linear Regression models rest on several assumptions regarding the relationship between the target variables and its predictors. For instance, regarding the errors, ε_t :

- they have mean zero, if not, the forecasts would be systematically biased.
- they don't display any autocorrelation, otherwise the resulting forecasting would be inefficient and incomplete, not utilizing all the available data.
- they are unrelated to the target variable, if not, there is information that should be included in the systematic part of the model.

2.4.2.2 Random Forest and XGBoost: Overview and Key Concepts

Random Forest consists of an ensemble machine learning method that can be used for both classification and regression tasks. Ensemble machine learning methods aggregate two or more learners (models) in order to produce more accurate prediction (Zhou, 2012).

Although initially developed by Tin Kam Ho (Tin Kam Ho, 1998), random forest was registered as a trademark by Leo Breiman and Adele Cutler (Breiman, 2001). The method combines the output of multiple decision trees to calculate a single result. In the case of regression tasks, the output of a random forest model is the average of the predictions of the trees.; For classification tasks, it typically involves majority voting.

As mentioned, Random Forest consists of averaging the output of multiple decision trees to produce the final output of the model. Therefore, it is important to briefly describe the concept of Decision Tree. A decision tree is a non-parametric supervised learning algorithm, which can be used for both classification and regression tasks. It consists of a hierarchical, tree structure with a root node, branches, internal nodes and leaf nodes (IBM, 2025a).

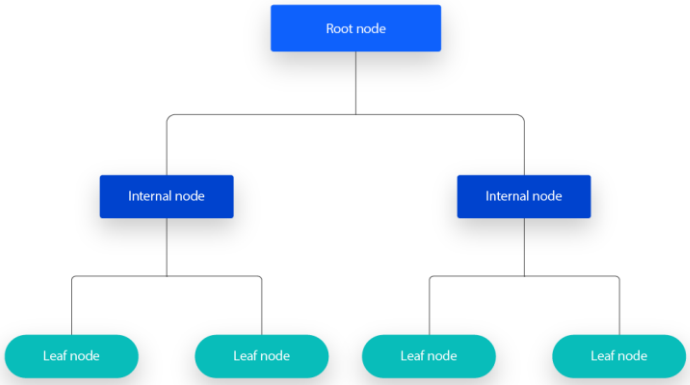

Figure 8 - Decision Tree Structure. Source:(IBM, 2025a)

Decision Tree learning uses a greedy, top-down divide-and-conquer approach to find the best split points and classify data. There is a tendency of larger trees to struggle with data fragmentation, where too few data points fall into each subtree. This fragmentation can lead to the risk of overfitting (IBM, 2025a).

Random Forest addresses this issue by extending the **bagging** (bootstrap aggregating) technique proposed by **Leo Breiman** (1996). In bagging, multiple models are trained on random samples (with replacement) of the original dataset. Random Forest goes a step further by introducing **feature randomness**: rather than considering all features at each split (as a decision tree does), it selects a **random subset** of features. This encourages **decorrelation** between trees, which improves generalization performance (IBM, 2025b).

Random forest model has three main hyperparameters: node size, number of trees and the number of features sampled.

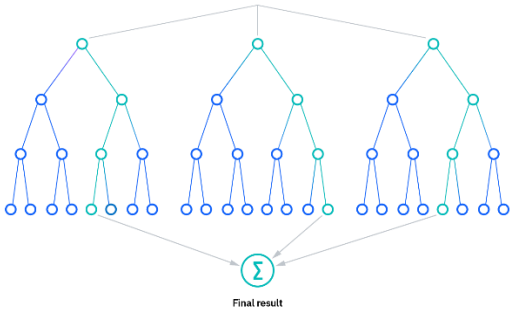

Figure 9 - Random Forest. Source: (IBM, 2025b)

XGBoost (eXtreme Gradient Boosting) refers to a distributed, open-source, machine learning library that implements gradient boosted decision trees.

Unlike Random Forest, which builds trees independently and in parallel, gradient boosting builds trees sequentially. It begins with a weak learner to make predictions and iteratively improve performance by building new trees that correct the residual errors of the previous ones.

At each step, the algorithm tries to minimize a loss function, such as squared error, using gradient descent to improve the optimization process (IBM, 2025c).

XGBoost distinguishes itself from other gradient boosting implementations with several advanced features (xgboost developers, 2022):

- **Parallel and distributed computing**: the library stores data in in-memory units called blocks. These blocks can be distributed separately across different machines.
- **Cache-aware prefetching algorithm**: this allows the library to offer better performance compared to other implementations of the gradient boosting algorithm.
- **Built in regularization**: extends regular gradient boosting, by including regularization as part of the learning objective.
- **Handling missing values**: uses a sparsity-aware algorithm for sparse data.

3 State-of-the-Art Review

This chapter outlines the state-of-the-art review conducted within the scope of this project. By addressing the predefined research question and detailing the research methodology employed, this chapter aims to provide comprehensive insights into the latest developments, technologies, and trends in the field of machine learning approaches for forecasting economic recessions. Additionally, it highlights the main challenges currently faced in this domain.

3.1 Research Questions

In the context of the state-of-the-art review conducted, the following primary research question was formulated:

- **RQ1**: What are the key techniques and economic indicators used in machine learning approaches to predict economic recessions?

The research question was framed using the PICOCS model (Population, Intervention, Comparison, Outcomes, Context, Study Design). Table 1 provides mapping of the research question to the elements of the PICOCS framework.

Table 1 - RQ1 framed using PICOCS model

	RQ1
Population	Economic Data/Research Studies
Intervention	Machine Learning approaches
Comparison	Not Applicable
Outcomes	Identification of key economic indicators and effective techniques for recession prediction
Context	Economic environments
Study Design	Empirical studies

3.2 Research Methodology

To provide answers to the research question defined, a Systematic Mapping Review was carried out, using the PRISMA methodology.

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology is described in detail in the PRISMA 2020 statement paper, consisting of a 27-item checklist and a flow diagram (Page *et al.*, 2021). Although designed primarily for systematic review of studies that evaluate the effects of health interventions, PRISMA checklist are applicable to systematic reviews evaluating other interventions. According to the authors of the statement paper, PRISMA 2020 is "intended for use in systematic reviews that include synthesis (such as pairwise meta-analysis or other statistical synthesis methods) or do not include synthesis (for example, because only one eligible study is identified)".

3.2.1 Data Sources

The data sources present in Table 2 were used to search and retrieve the documents in the context of this research.

Table 2 – Data Sources

Identifier	Database	URL
DS1	Web of Science – Core Collection	Document Search - Web of Science Core Collection
DS2	B-ON	b-on

3.2.2 Search Terms

Once the main research question was established, relevant search terms were identified to query the data sources. A search string, present in Code Snippet 1, was constructed to target the fields *Title*, *Abstract*, and *Keywords*, using the "OR" operator to combine these fields effectively.

```
("economic recession" OR "economic downturn" OR "economic crisis" OR
"recession" OR "financial crisis") AND ("machine learning" OR "deep learning"
OR "artificial intelligence" OR "ML models" OR "neural networks" OR "ML" OR
"DL") AND ("prediction" OR "forecasting") AND ("EU" OR "Europe" OR "US" OR
"United States" OR "United Kingdom" OR "Eurozone" OR "Portugal")
```

Code Snippet 1 - Search String

The search string was designed to balance precision and recall by broadly including diverse machine learning techniques, such as "machine learning," "deep learning," and "neural networks," while focusing on the specific prediction of recessions. Targeted terms like "economic recession" and "financial crisis" were used alongside geographic filters for regions

like the EU, US, and Portugal. This approach ensured comprehensive coverage of relevant studies while minimizing irrelevant results.

3.2.3 Eligibility Criteria

In accordance with the PRISMA guidelines (Page *et al.*, 2021), the following set of inclusion and exclusion criteria were defined for the screening and eligibility phases of the research process.

- Inclusion criteria:
  - **IC1**: The study must apply machine learning techniques.
  - **IC2**: The study must specifically focus on macroeconomic forecasting of recessions.
  - **IC3**: The study must be written in English.
  - **IC4**: The study must have been published within the last 10 years (2014-2024)

- Exclusion criteria:
  - **EC1**: The study does not specifically incorporate machine learning techniques.
  - **EC2**: The study was published before 2014.
  - **EC3**: The study is not written in English.
  - **EC4**: The study does not include experimental results or performance evaluations.
  - **EC5**: The study does not focus on the forecast of economic recessions.
  - **EC6**: The study is only presented as poster, abstract, or conference proceedings, without full-text availability.

A document was included if it met all the inclusion criteria. Conversely, it was excluded if it fulfilled at least one of the exclusion criteria.

3.2.4 Collection Process

As outlined earlier, the process of collecting and reviewing the studies included in this state-of-the-art research adhered to the PRISMA methodology. The structured approach, described by the PRISMA guidelines, consists of four distinct phases (Page *et al.*, 2021). The first phase, **Identification**, involves conducting a comprehensive search across multiple data sources to ensure broad coverage of the research topic by identifying all potentially relevant studies. During this phase, duplicate records are detected and removed. In the **Screening** phase, the title and abstracts of studies identified are reviewed against the inclusion and exclusion criteria previously defined. The studies that clearly do not meet the inclusion criteria or meet any exclusion criteria are excluded at this stage. After this stage, in the **Eligibility**, the full-text articles are retrieved and evaluated in-depth against the inclusion and exclusion criteria, with any study that fails to meet the criteria being excluded from the review. In the final phase, **Included,** the final set of studies is included for further extraction of data and analysis. In the

PRISMA flow diagram shown in Figure 10, the multiple phases involved in this systematic review are represented.

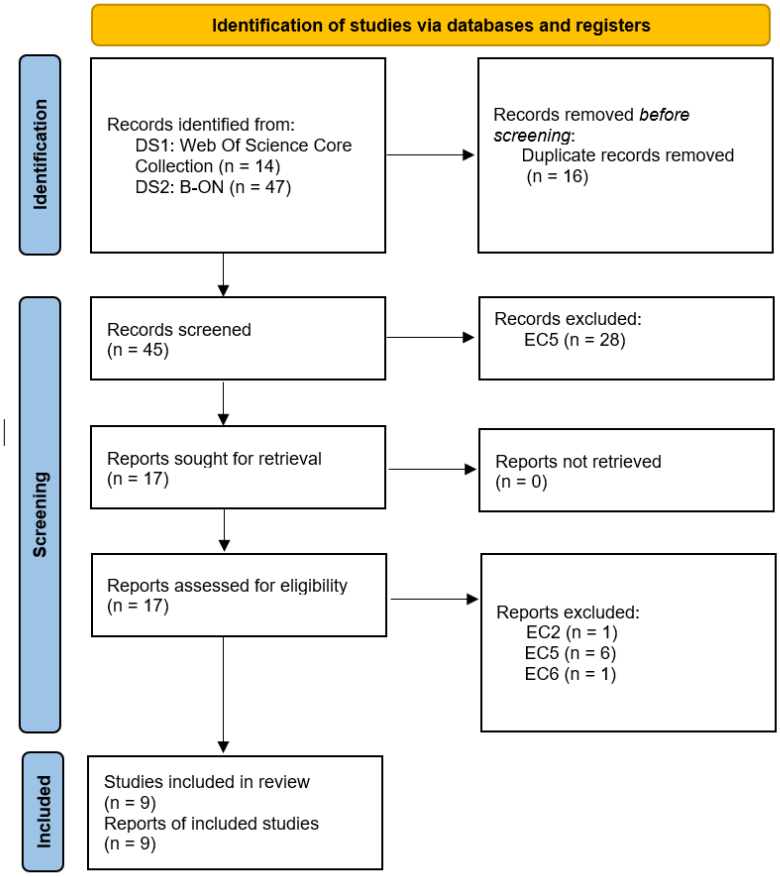

Figure 10 - PRISMA Flow Diagram

As illustrated in Figure 10, during the collection process for this systematic review, a total of 61 studies were retrieved in the Identification phase from the defined data sources, with 16 duplicate records removed. In the Screening phase, 45 studies were analysed, with 28 being excluded due to meeting EC5. Finally, the full-text retrieval and in-depth analysis was conducted on 17 studies, of which 8 were excluded from the final list based on previously defined exclusion criteria (1 for EC2, 6 for EC5, and 1 for EC6).

3.3 Results

A total of nine studies were included in the review. A comprehensive full-text analysis of these papers was performed to extract and categorize relevant data addressing the defined research question. The results of this process are presented in Table 3, where the extracted data from each study are organized into the following categories:

- **Source:** reference for the study.
- **Pub. Date:** year of publication.
- **Data Origin:** geographic location(s) or region(s) of the dataset.
- **Time Period:** the time spans used for training and testing the models.
- **Economic Indicators:** economic variables used in the study.
- **ML Techniques:** machine learning methods or algorithms applied.
- **Results:** concise summary of the findings, such as model performance or key conclusions.

Table 3 - Mapping of Results Extracted from Economic Recession Prediction Studies

Source	Pub. Date	Data Origin	Time Period	Economic Indicators	ML Techniques	Results
(Nyman and Ormerod, 2017)	2017	US, UK	Training: 1970Q2–1990Q1 (initial regression); extended incrementally to 1970Q2–2016Q1. Testing: Predicting sequentially from 1990Q2 to 2016Q2.	US: 3-Month Treasury Bill Rate; Yield on 10-Year US Government Bonds; Quarterly % change in the S&P 500; UK: 3-Month Treasury Bill Rate; Yield on 10-Year UK Government Bonds; Quarterly % change in the FTSE All Share Index; Ratio of Private Sector Debt to Current Price GDP	OLS, Random Forest	Random Forest outperforms OLS for 3 and 6-step ahead forecasts in both UK and US data predictions;
(Sties, 2017)	2017	US, Canada		US: 135 indicators across eight groups: Output & Income, Labour Market, Consumption, Orders & Sales, Money & Credit, Interest & Exchange Rates, Prices, and Stock Market.	Lasso Regression, Boosting Decision Trees	Lasso and Boosting outperform naive best subset selection. Canadian yield spreads are key predictors across all

				Canada: 445 monthly economic variables		forecast horizons.
(Nyman and Ormerod, 2020)	2020	US	Training: 1970Q2–1989Q2 (initial regression); extended incrementally to 1970Q2–2010Q3. Testing: Predicting sequentially from 1990Q3 to 2010Q4.	3 month Treasury Bill rate; Yield on 10 year US Government bonds; Quarterly percentage in the index of total share prices in the FRED database; Ratio of household debt to GDP; Ratio of non-financial corporate debt to GDP	Random Forest	RF predicts 2009 recession 18 months early (-2.54% for 2009Q1). Misses 1990/91 and 2001 recessions but flags <1% growth periods. High prediction variance noted.
(Zyatkov and Krivorotko, 2021a)	2021	US	Training: 1955–2020, divided into time windows for training, validation and testing; Testing: From 1976 to 2020, with six test samples (green blocks in Fig. 2). The model predictions for each test sample were used to form a time series describing the likelihood of a recession in the US economy over the next 6/12/24 months.	Effective Federal Funds Rate; Consumer Price Index (CPI); Gold price per ounce; 10-Year Treasury Rate; US Yield Curve; S&P 500 Index; Nonfarm Payrolls; Purchasing Manager's Index; U.Michigan Consumer Sentiment Index;	Logistic Regression, KNN, SVM, Random Forest, Gradient Boosting, FCNNs, LSTM	A FCNN with 9 input neurons, 8 hidden neurons (ReLU), and 2 output neurons (softmax) provided the most accurate recession predictions, minimizing errors better than other methods.
(Qilu, 2022)	2022	Jordà-Schularick-Taylor Macrohistory database (17 advanced economies)	Not specified	GDP; GDP per Capita; CPI; Money; Consumption; Investment; Credit; Yield Curve; Public Debt; Debt service ratio; Current account; Export; USD exchange rate; Global yield curve; Global credit;	Logit (as benchmark), XGBoost and Random Forest	XGBoost and random forest achieve AUROC values over 0.99, significantly outperforming the logit model (0.724). The removal of GDP improves AUROC in both ML models, suggesting it should be excluded. XGBoost and random forest show nearly

						identical predictive performance.
(Petropoulos *et al.*, 2023)	2023	US	Training: January 1973 to December 2005 (65% of the dataset) Testing: January 2006 to December 2018 (35% of the dataset), excluding the 2008-2010 crisis period to assess model stability through the economic cycle.	GDP; Government Debt % GDP; Real Estates Prices Yearly Growth; Yearly change in Unemployment Rate; CPI yearly growth; 10 year Gov bond yield changes; Government Expenses % GDP; Exports yearly growth; Annual return S&P 500;	BMA, MXNET, Bayesian ReLU and Bayesian LWTA	The study compares multiple models for macroeconomic forecasting, finding that Deep Learning models, especially Bayesian LWTA, outperform Bayesian Model Average (BMA). The Bayesian LWTA showed superior accuracy in forecasting key variables like bond yields, real estate prices, and stock market growth during the Subprime crisis, with lower MSE and MAE.
(Chung, 2023)	2023	US	The study examines predictions starting from November 2006, across five windows: nowcasting (current month), immediate-term (1 month ahead), short-term (3 months ahead), medium-term (6 months ahead), and long-term (12 months ahead), converting probability predictions into binary indicators for recession or boom.	24 different economic predictors that are grouped into 7 categories: Income, Labor Market, Money and Credit, Output, Financial Market, Housing Market, Prices	Logit, Ridge (statistical methods), FFNN, LSTM and GRU	GRU and LSTM outperform logit models across all forecast horizons, with GRU leading in nowcasting, short-term, and long-term setups, and LSTM excelling in medium-term forecasts. Both neural networks handle class imbalance well, with GRU achieving the

						best overall performance.
(Wang *et al.*, 2024)	2024	G7, BRICS, Australia	The training data set for Russia spans from 1991 to 2010, while the test data set runs from 2011 to 2019. For other countries, the training data set covers the period from 1980 to 2010, with the test data set from 2011 to 2019.	GDP; Net population; CPI; Employment rate; Share of labour compensation in GDP at current national prices; Exchange rate; GDP per-capita at current; Purchasing power parity (PPP); Price levels (macroeconomics perspectives: production, expenditure, and trade); Finance; Human resources;	LSTM, BD-LSTM, ED-LSTM and CNNs	Shuffling the training data effectively prevented model overfitting. Comparing traditional models (ARIMA, VAR) with deep learning models (CNN, LSTM, BD-LSTM, ED-LSTM) showed that ED-LSTM performed the best.
(Pontes *et al.*, 2024)	2024	US, Eurozone	Eurozone: Train: 1981-2001 Validation: 2002-2011 Test: 2012-2022 US: Train: 1969-2001 Validation: 2000-2009 Test: 2010-2022	GDP, inflation rate, industrial performance, and market indices (e.g., S&P 500, Dow Jones, and commodities like oil and gold)	Multinomial Logistic Regression, SVM and Multi-layer Perceptron (MLP ANN)	Eurozone: MLP ANN achieved best results with 55.81 % F-Score. In US, Multinomial Logistic Regression achieved the best results with 76.92 % F-Score.

3.4 Discussion

RQ1: What are the key techniques and economic indicators used in machine learning approaches to predict economic recessions?

An analysis of the results presented in Table 3 reveals that, over the past decade of research, a diverse array of machine learning techniques has been applied to predict economic recessions. Although the sample size of analysed papers is relatively small, a clear pattern emerges in the use of these techniques. For instance, early studies predominantly utilized traditional machine learning approaches such as Ordinary Least Squares and Lasso Regressions, as well as Random Forest or Boosting Decision Trees (Nyman and Ormerod, 2017, 2020; Sties, 2017). However, since 2021, there has been a noticeable shift toward the usage of Deep Learning techniques,

including Feed Forward Neural Networks and Recurrent Neural Networks, such as Long short-term memory and Gated Recurrent Unit. (Chung, 2023; Petropoulos *et al.*, 2023; Zyatkov and Krivorotko, 2021b). In terms of effectiveness of the approaches, deep learning models generally outperform traditional models in predicting economic recessions. For instance, in (Chung, 2023), GRU and LSTM models outperformed a Logistic Regression model across various forecast horizons. This trend is also evident, for instance, in the studies performed by (Zyatkov and Krivorotko, 2021b) and (Wang *et al.*, 2024), where Fully Connected Neural Networks (FCCN), LSTM and ED-LSTM were the best performing models.

The success of machine learning techniques in predicting economic recessions is closely linked to the selection of key economic indicators. Commonly used indicators across studies include GDP, CPI, interest rates, and yield curves. Analysing the influence of these economic indicators on recession prediction models is also crucial, as it offers valuable insights into their effectiveness and the relative importance of each indicator in forecasting economic downturns. For instance, in a study by (Qilu, 2022), which analysed economic data from 17 advanced economies, a Shapley summary plot was used to assess the impact of various economic indicators on the XGBoost model's predictions. Figure 11 illustrates the Shapley summary plot. Qilu concludes that the two most influential variables in the model are the global and domestic yield curves, both of which had a considerable impact, followed by consumption and CPI.

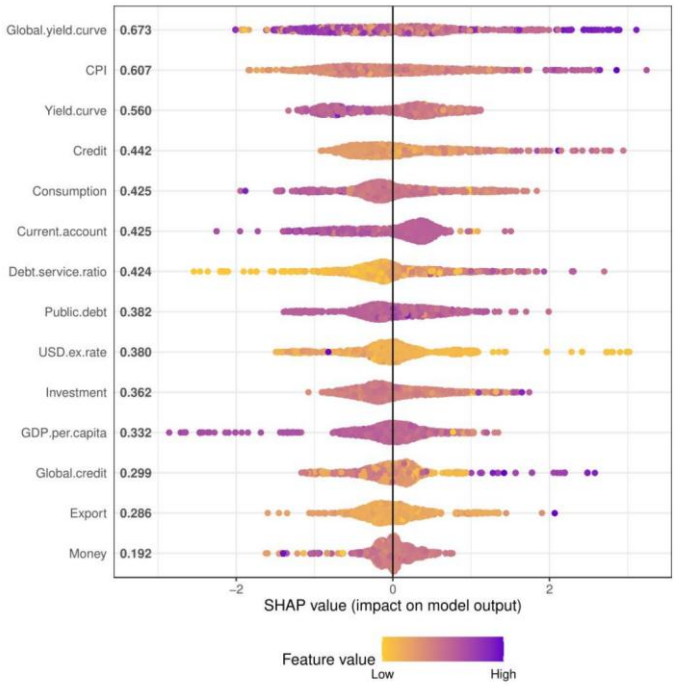

Figure 11 - Economic Indicators importance in XGBoost model. Source: (Qilu, 2022)

When it comes to the training and testing methodology, a recurring pattern emerges across the studies. Many studies adopt rolling or expanding windows to simulate real-world forecasting scenarios, where models are trained on historical data and tested on progressively newer data. This approach allows for a dynamic assessment of model performance as new economic events unfold. For example, (Nyman and Ormerod, 2017) employed a rolling window approach, where regressions were initially performed using data from 1970Q2 to 1990Q1 to predict 1990Q2. The training period was then incrementally extended by one quarter until the final regression included data from 1970Q2 to 2016Q1 to predict 2016Q2. Different time windows were also explored and tested in the different studies. For instance, in the paper by (Chung, 2023), five different time windows were tested: nowcasting (current month), immediate-term (1 month ahead), short-term (3 months ahead), medium-term (6 months ahead) and long-term (12 months ahead). In this study It is highlighted how different models, particularly GRU and LSTM performed better across various prediction horizons. This indicates that the model's predictive accuracy may be influenced by the specific forecasting window.

3.5 Challenges and Limitations

One common challenge noted across studies is the risk of overfitting, especially with complex models such as deep learning techniques. The study by (Wang *et al.*, 2024) demonstrates how shuffling training data can help mitigate overfitting. Additionally, the variability in model performance across studies, evidenced by the missed prediction of the 1990/91 and 2001 recessions in the study conducted by (Nyman and Ormerod, 2020), suggests that future research could focus on improving model robustness and generalization to avoid performance degradation in unseen data. Finally, (Qilu, 2022) highlights a notable limitation in recession prediction: the inherent imbalance of the target variable, with crisis periods being significantly underrepresented compared to non-crisis periods

3.6 Research Gaps

Despite significant advancements in applying machine learning techniques to predict economic recessions, the state-of-the-art review highlights research gaps that warrant further exploration. One notable limitation is the lack of regional diversity in the datasets used. Most studies concentrate on large, advanced economies such as the US, UK, or Canada. While some research extends to the broader European economy, there is a noticeable scarcity of studies examining recession prediction in mid-sized economies, such as Portugal.

4 Exploratory Data Analysis

One of the most important phases in the life cycle of a data mining project, according to the CRISP-DM methodology, is the **"Data Understanding"** phase. As described by (Chapman *et al.*, 2000), this phase "starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information". This chapter aims to describe the process of **Exploratory Data Analysis (EDA)** within the scope of this project. It begins by introducing the dataset used, followed by an overview of statistical summaries, visual patterns, and correlation analysis. The final subchapter summarizes the key findings from the EDA process.

4.1 Dataset overview

The dataset used in the context of this project comprises 244 macroeconomic numerical timeseries as explanatory variables and one target variable, Real Gross Domestic Product (GDPC1). These series represent key indicators of the US economy and are grouped in 14 categories, including metrics related to output, labour markets, consumption, investment, interest rates and prices. The data was retrieved from the Federal Reserve's FRED-QD database, a quarterly frequency macroeconomic database (McCracken and Ng, 2020). As mentioned, all the feature variables consist of timeseries recorded at a quarterly frequency and span the period from 1st quarter (Q1) 1959 to 3rd quarter (Q3) 2024, resulting in a total of 263 observations per series. In addition to the observations, the FRED-QD database also includes a tcode value for each time series, indicating the type of transformation needed to make the series stationary:

- (1): no transformation required.
- (2): first difference.
- (3): second difference.

- (4): $\log(x_t)$.
- (5): first difference of $\log(x_t)$.
- (6): second difference of $\log(x_t)$.
- (7): growth rate approximation $\boldsymbol{\Delta(x_t / x_{t-1} - 1.0)}$

The FRED-QD database, in addition to originating from a credible source in US economy statistics, it has also been used in multiple prior studies (Chung, 2023) (Sties, 2017)(McCracken and Ng, 2020), making it suitable within the scope of this project.

4.2 Data Quality Analysis

An initial analysis of the dataset's quality was conducted, focusing on the presence of missing values across indicators. Figure 12 displays the top 30 features with the highest proportion of missing observations in the dataset. Among them, variables such as "**TWEXAFEGSMTHx**" (Trade Weighted U.S. Dollar Index against major currencies, monthly, from the Federal Reserve Board), "**SPCS20RSA**" (S&P/Case-Shiller 20-City Composite Home Price Index, seasonally adjusted) and "**EXUEU**" (U.S. Dollar to Euro Spot Exchange Rate) show some of the largest shares of missing data.

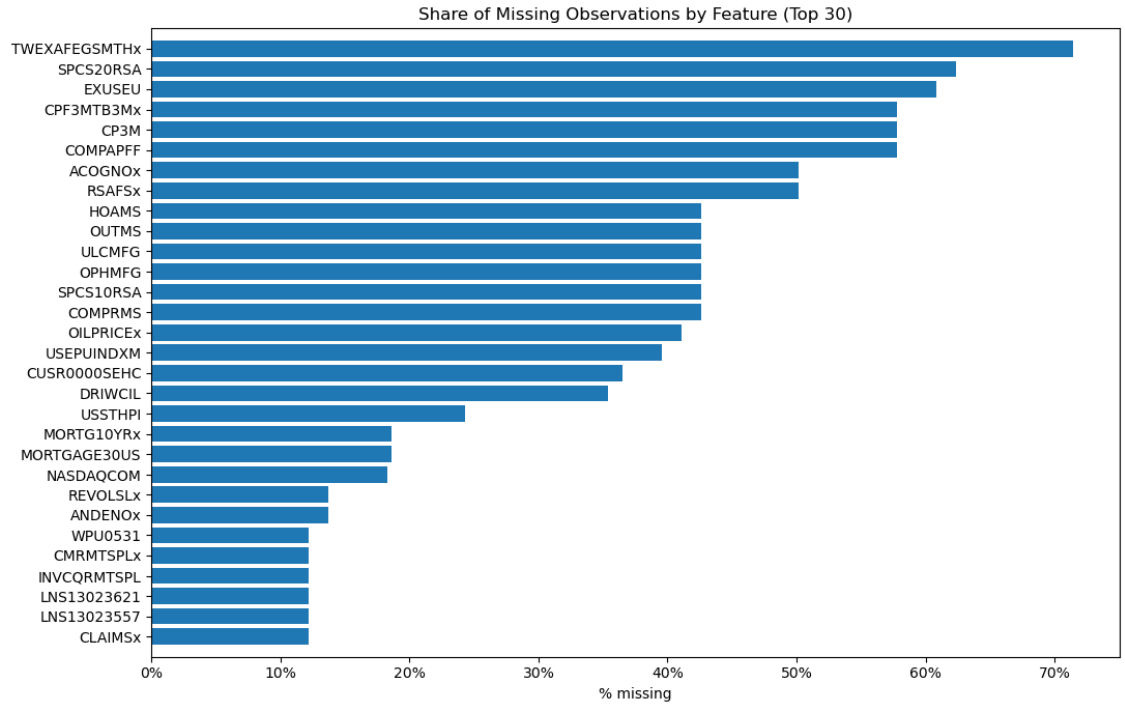

Figure 12 - Share of Missing Observations by Feature (Top 30)

Most of the features with missing observations are explained not due to errors in retrieval, but rather due to the scope and release periods of the data providers. For instance, the data

coverage of variable "**TWEXAFEGSMTHx"** only starts in January 2006 (Board of Governors of the Federal Reserve System (US), 2025a) and in the case of variable "**EXUEU",** the Euro was only introduced in 1999, so the data coverage only starts in January 1999 (Board of Governors of the Federal Reserve System (US), 2025b)**.**

4.3 Target Variable Analysis – Time Series Visualization

As introduced in subchapter 2.1, the problem is framed as a regression task where the target variable is the **quarterly real GDP,** identified in the FRED dataset with column "GDPC1". In this subchapter, the focus will be on exploring the behaviour of this variable over time.

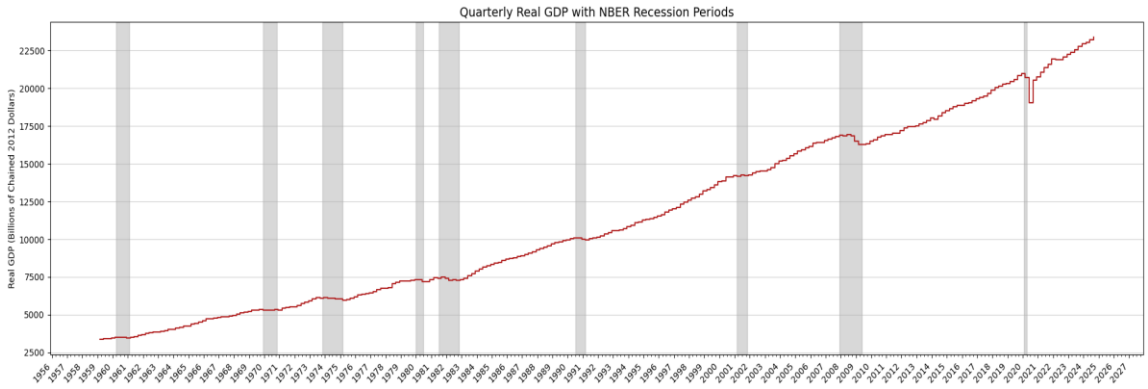

Figure 13 - Quarterly Real GDP with NBER Recession Periods

As presented in Figure 13, we can observe the time series visualization of the target variable, quarterly real GDP. Overall, the series shows a consistent upward trend, reflecting the long-term growth of the US economy, with some temporary contractions periods that align with major recession periods.

The shaded areas of the plot represent the recession periods, as defined by the NBER. These periods align closely with visible interruptions in the growth path of GDP, illustrating how recession periods, although temporary, are significant deviations from the long-run trend.

To better understand the cyclical dynamics of the time series, we can observe in Figure 14, two common measures used in econometrics: **quarter-on-quarter** and **year-on-year** growth rates, which reflects short and long-term fluctuations respectively.

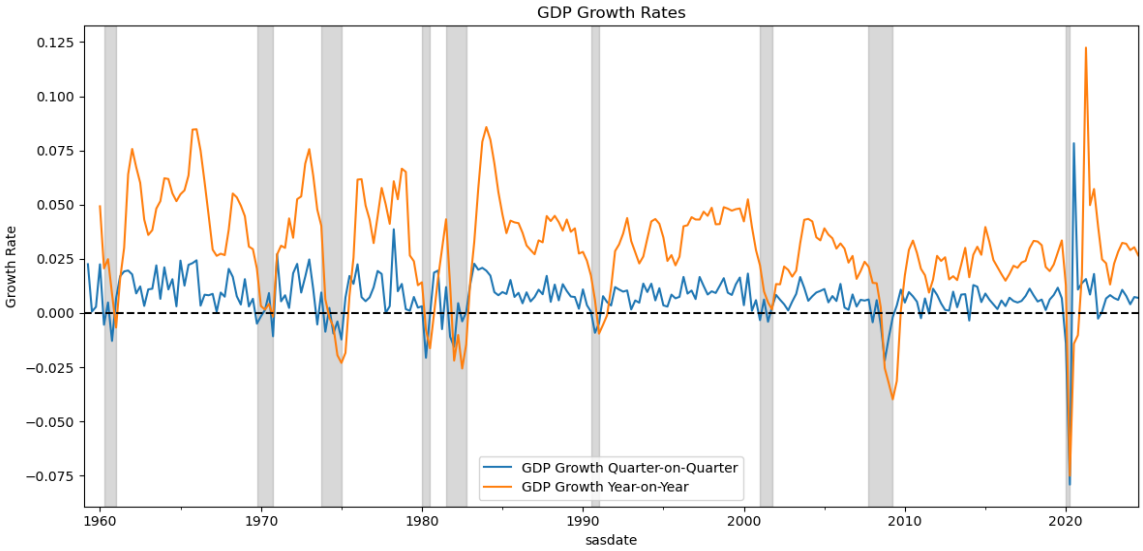

Figure 14 - GDP Growth Rates

This graph makes the recessions periods stand out more clearly, with negative growth rates coinciding with the official NBER recession periods (represented in the shaded areas). The negative growth rates are particularly noticeable in the early 1980s double-dip recession, the 2007-2009 Global Financial Crisis, and the 2020 COVID-19 crisis.

To examine more in depth the statistical properties of GDP, an **STL decomposition** was applied to the time series. As previously described in subchapter 2.2.1, a time series can be decomposed in trend, seasonal and residual components. In Figure 15 we can observe these components for the time series being studied.

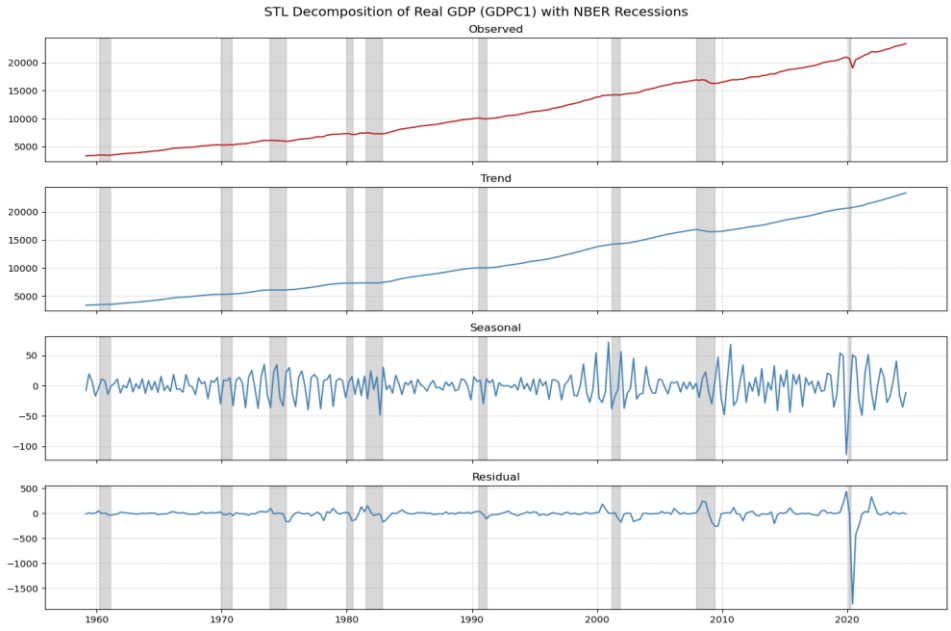

Figure 15 - STL Decomposition of Real GDP (GDPC1) with NBER Recessions

4.4 Initial Feature Screening

Given the high dimensionality of the initial dataset, with around 244 numerical macroeconomic timeseries, an initial feature screening process was performed. The objective of this process was to focus the analysis on the features with the strongest relationship to the target variable. For this purpose, the **Pearson Correlation Coefficient** was computed between each explanatory variable and the target variable.

In the scope of this project, features were retained only if they exhibited both a large absolute correlation coefficient ($|r| > 0{,}90$), indicating a strong linear association with the target variable, and statistical significance ($\rho < 0{,}05$), as determined by the p-value test.

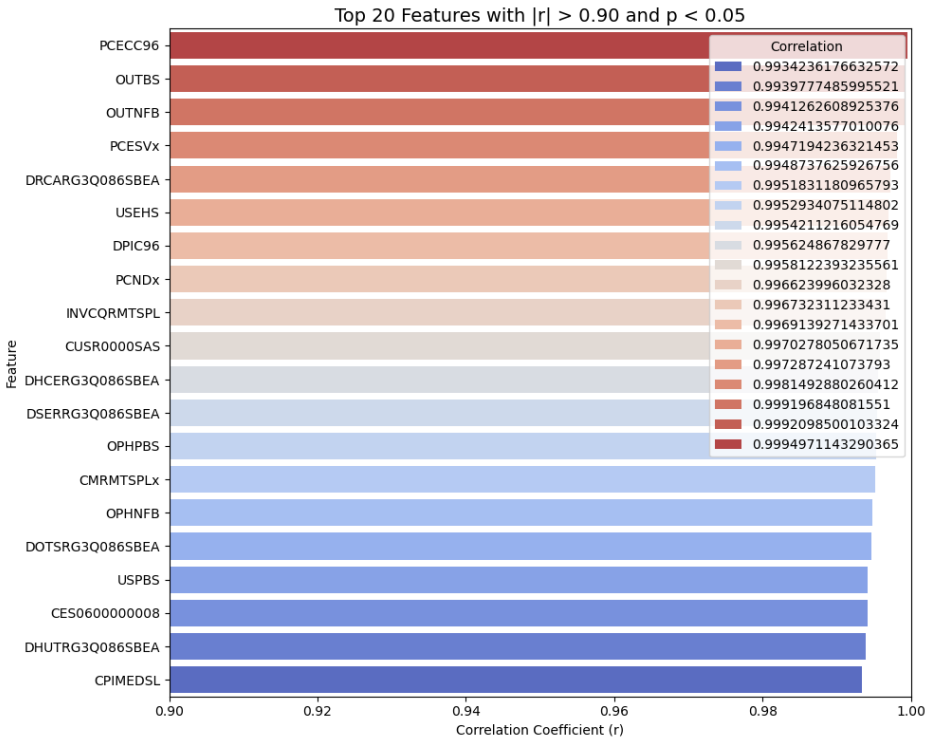

Figure 16 - Pearson Correlation results

The results presented Figure 16 show the top 20 features with the strongest linear association with the target variable, as measured by the Pearson Correlation coefficient.

By applying the predefined thresholds ($|r| > 0{,}90$, $\rho < 0{,}05$), the dimensionality of the dataset was reduced substantially, from 244 to **125** continuous numerical variables.

4.5 Features Behaviour and Relationships

After the initial feature screening process, a large number of variables (125) were still present in the dataset. As such, individually analysing the relationship between each feature variable and the target variable, although possible, it is not a practical solution for the scope of this dissertation and would add little analytical clarity for the objective of this subsection. The original FRED-QD dataset appendix organizes the dataset into fourteen broad economic categories such as: *NIPA*, *Employment and Unemployment*, *Housing*, etc (McCracken and Ng, 2020) . To simplify the scope of the analysis, a representative feature was selected from each FRED-QD group to illustrate the typical behaviour of that category in relation to the target variable. It should be noted that the categories *Stock Markets* and *Other* were not represented in the screened subset of 125 features and were therefore excluded from this analysis.

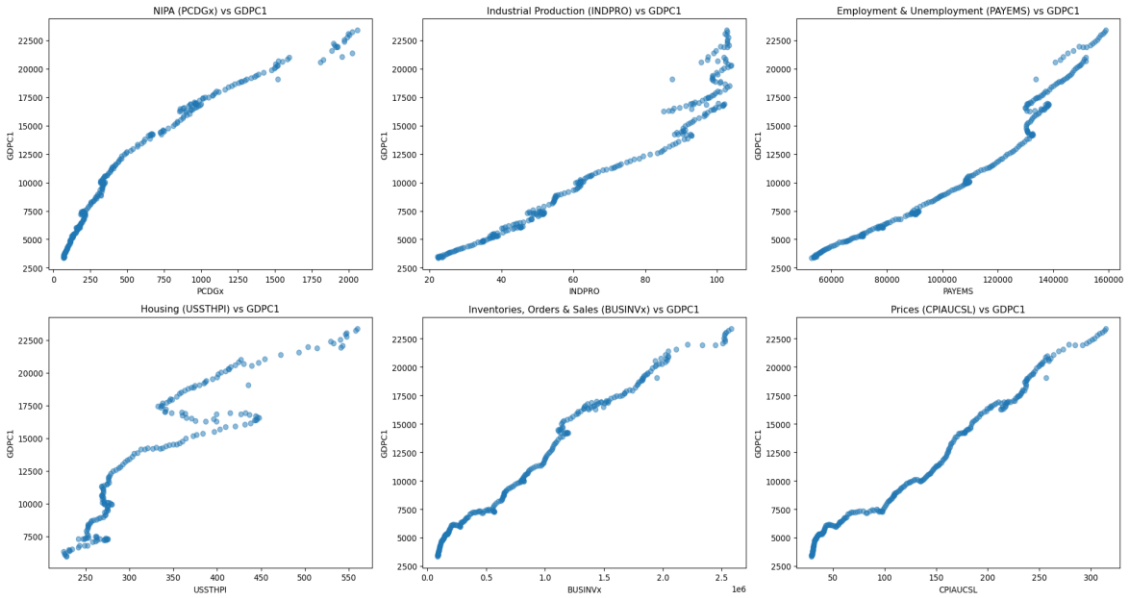

Figure 17 - Scatterplots between representative features of each FRED-QD group and the target variable (GDPC1), Part 1

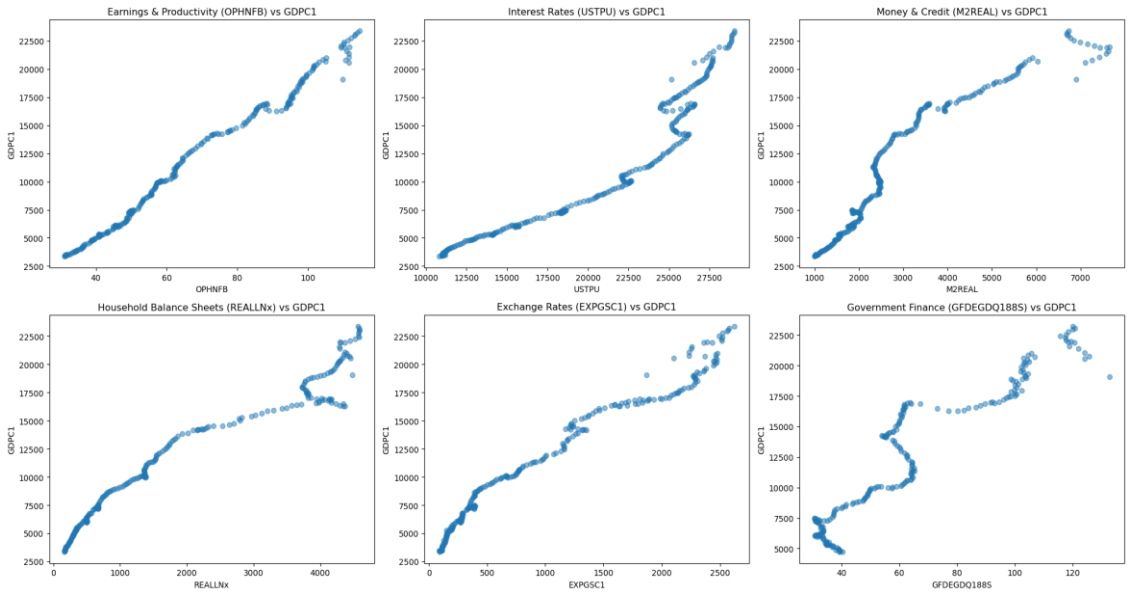

Figure 18 - Scatterplots between representative features of each FRED-QD group and the target variable (GDPC1), Part 2

The scatterplots presented in Figure 17 and Figure 18 display the relationship between real GDP (GDPC1) and one representative feature from each of the twelve FRED-QD categories retained after screening. By carefully analysing the plots, we can extract two important insights. Firstly, real activity indicators such as industrial production, employment and prices seem to have a linear association with GDP, confirming their importance to the overall output of the US economy. Secondly, features associated with categories such as *Government Finance*, *Interest Rates* and *Money & Credit*, seem to display more complex and inconsistent patterns, including nonlinearities, periods of divergence and structural breaks.

While scatterplots provide an overview of the long-run association between two variables, they don't capture whether these relationships remain stable over time or vary across different phases of the series. To provide a more accurate overview of the relationship between GDP and each representative feature of the twelve FRED-QD categories over time, **rolling correlation plots** were developed. These plots allow the correlation coefficient to be calculated repeatedly over a moving window, this way creating a dynamic perspective on how the strength and direction of the association evolves over time.

For this analysis, a 20 quarter (five-year) rolling window was used, meaning that at each point in time the correlation coefficient is calculated using the most recent five years of data.

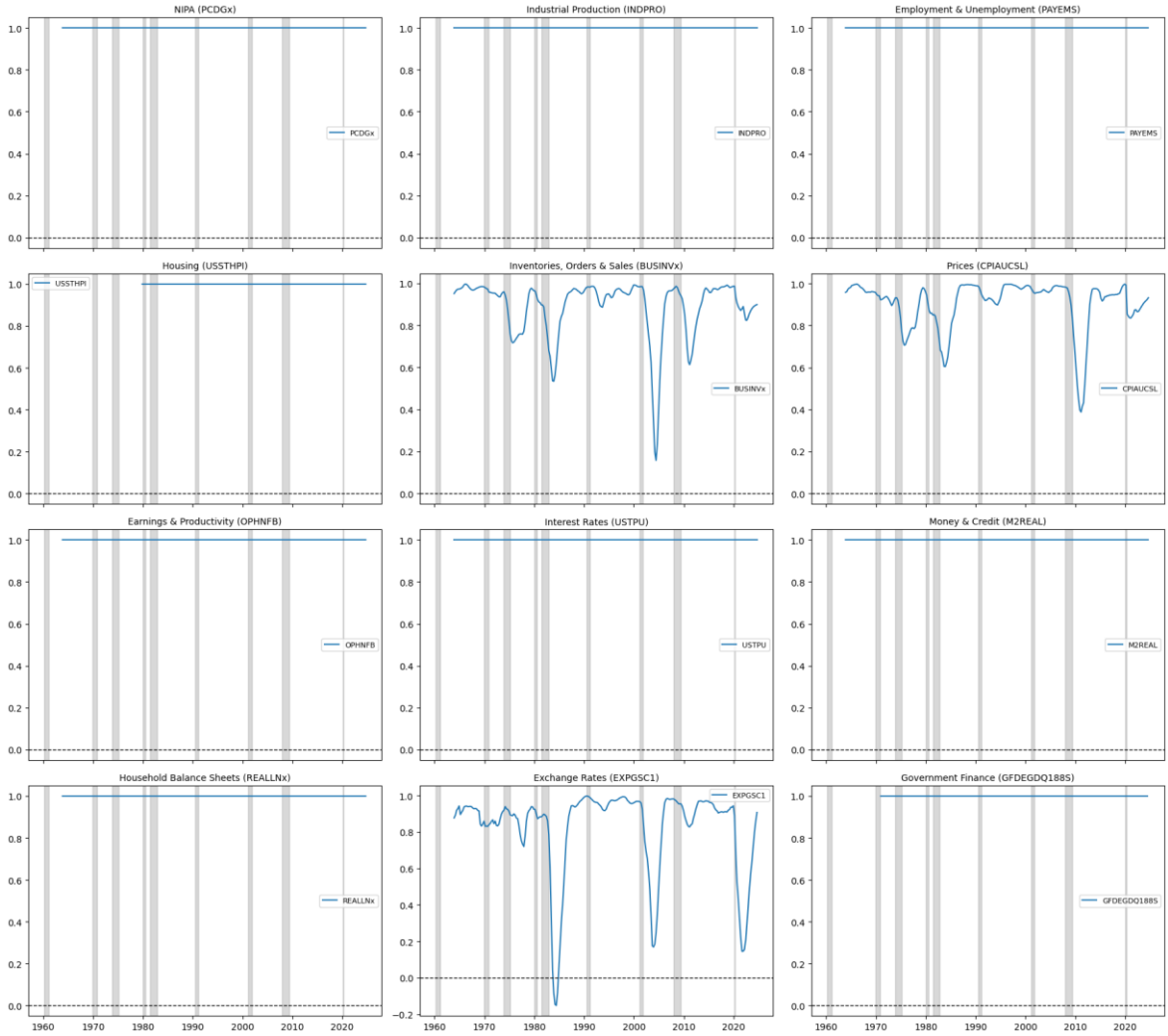

Figure 19 - Rolling correlations (20-quarter window) between GDP and representative features of the FRED-QD categories, with shaded areas indicating NBER recession periods.

Analysing Figure 19, we can observe two distinct patterns. In several categories, such as: *NIPA*, *Industrial Production*, *Employment & Unemployment*, *Housing*, *Earnings & Productivity*, *Interest Rates*, Money & Credit, *Household Balance Sheets* and *Government Finance*, the rolling correlation coefficient remains close to one, reflecting a very strong relationship with GDP. By contrast, categories such as: *Inventories*, *Orders & Sales; Prices* and Exchange Rates, seem to display more inconsistent time-varying behaviour, with the correlations weaking during periods of recessions, as highlighted by the shaded areas, suggesting that their association with GDP is more sensitive to cyclical and structural shocks.

4.6 Stationarity – AD Fuller Test

As introduced in subchapter 2.2.1 (Characteristics of Time Series), **Stationarity** is the condition in which a time series exhibits, over different time periods, constant mean, constant variance or standard deviation and that the autocorrelation does not change over time (Hyndman and Athanasopoulos, 2021).

To test whether a series is stationary, one of the most widely adopted statistical tests is the Augmented Dickey-Fuller (ADF) test, where the null hypothesis states that the series contains a unit root and is therefore non-stationary. Rejecting the null hypothesis proves that the series is stationary.

The ADF Test was applied to the target variable of this study, quarterly real GDP (GDPC1).

Table 4 - ADF Test Results for GDPC1

ADF Statistic	p-value	1% Critical Value	5% Critical Value	10% Critical Value	Stationary?
2.7836	1.0000	-3.4557	-2.8727	-2.5727	No

The results presented in Table 4 indicate that the GDPC1 series is non-stationary, as the p-value is above conventional significance ($\rho > 0{,}05$) threshold.

After applying first differencing to the target GDP series, the results of the ADF test are presented in Table 5. The test statistic of -18.2796 is below the 1% critical value of -3.4557, and the associated p-value is zero. The results prove that we can reject the null hypothesis of a unit root, this way confirming that the differenced GDP series is stationarity.

Table 5 - ADF Test Results for GDPC1 with First Differencing

ADF Statistic	p-value	1% Critical Value	5% Critical Value	10% Critical Value	Stationary?
-18.2796	0.0000	-3.4557	-2.8727	-2.5727	Yes

4.7 PCA – Based Exploratory Analysis

Given the high dimensionality of the initial dataset considered in this project, comprising 244 explanatory variables from the FRED-QD database, PCA was used as a dimensionality reduction technique. This approach is consistent with the methodology applied in the original FRED-QD working paper that accompanies the dataset (McCracken and Ng, 2021), where PCA components were also used to summarize the economic indicators for forecasting purposes.

As PCA requires the use of normalized features and does not accept missing values, before applying PCA, missing values were handled by applying *forward fill* and *backward fill* imputation, ensuring continuity across time. In addition, all explanatory variables were normalized, so that differences in scale do not disproportionately affect the construction of components.

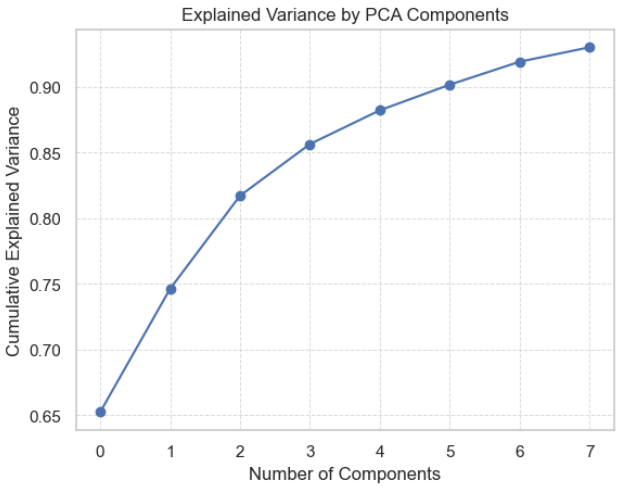

Figure 20 - Cumulative explained variance by PCA components.

As can be observed in Figure 20, the first three components explain more than 80% of the total variance, and seven components explain over 90%, indicating that the original 244 explanatory variables can be summarized in a small number of principal components.

After reducing the dimensionality of the explanatory variables through PCA and evaluating the cumulative explained variance to determine an appropriate number of components, the first eight principal components (PC1-PC8) were extracted and analysed against GDP target series (GDPC1). Figure 21 presents the scatterplots of each principal component against GDP.

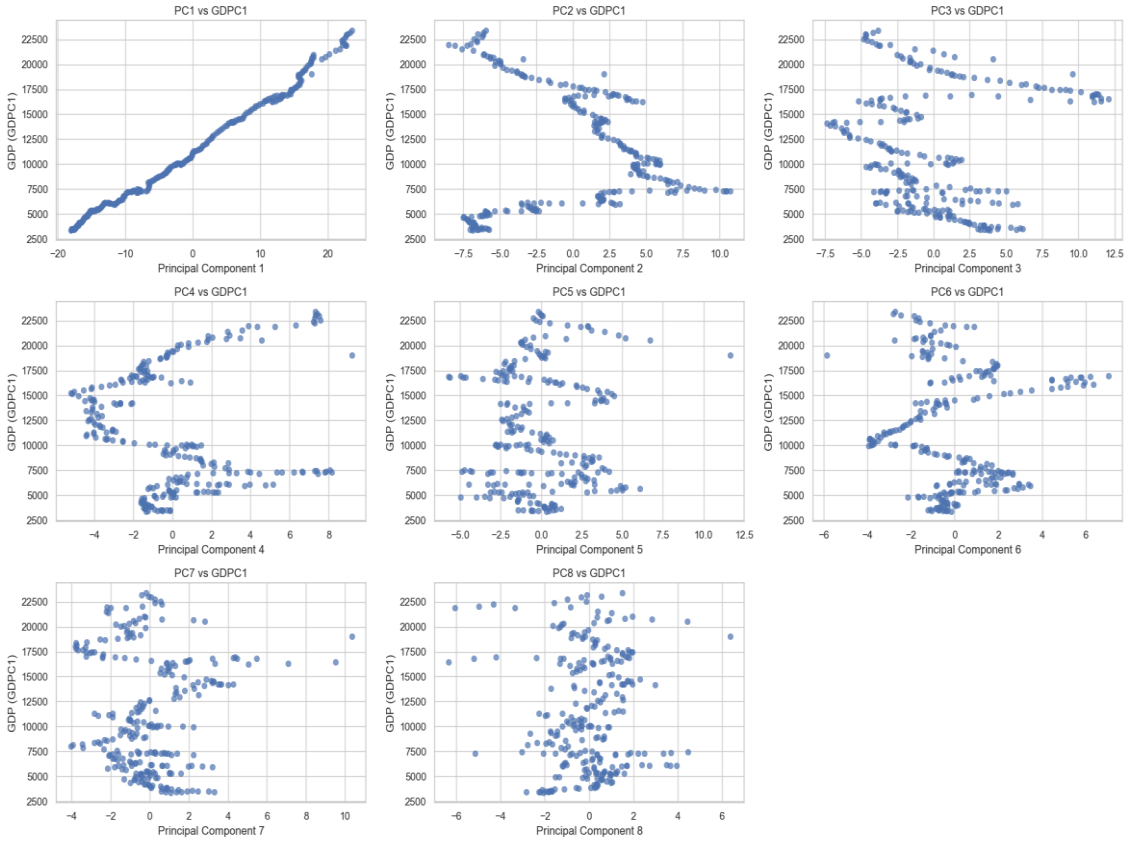

Figure 21 - Scatterplots of the first eight principal components (PC1–PC8) against GDP (GDPC1).

Analysing Figure 21, we can observe that the first principal component (PC1) displays a strong positive linear relationship with GDP. This is somewhat expected, since PC1 captures the direction of maximum variance in the dataset and has a cumulative explained variance of over 65%. The remaining components (PC2-PC8) seem to exhibit nonlinear and dispersed relationship with target GDP series, indicating that these components may be capturing cyclical fluctuations or structural breaks.

4.8 Key Findings

The EDA process is helpful in providing a comprehensive understanding of the structure, properties and suitability of the dataset in the context of forecasting the target variable (GDPC1). Across the seven subchapters of this section, the main insights and findings can be summarized as follows:

- **Dataset Characteristics**:
    - The dataset contains 244 explanatory continuous numerical variables and one target variable (real GDP, GDPC1), all recorded at a quarterly frequency between 1959Q1 and 2024Q3.
- **Data Quality**:
    - Missing values are present, particularly for variables introduced in later years. For example, the Euro exchange rate only begins in 1999.
- **Target Variable Properties (GDPC1):**
    - GDP shows a strong long-term upward trend with temporary contractions, corresponding to recession periods as defined by NBER.
    - STL decomposition confirms that the series shows a dominant trend, moderate cyclical fluctuations and residual irregularities.
    - Augmented Dickey-Fuller (ADF) test indicates that the series is non-stationary but becomes stationary after applying first differencing.
- **Feature Relationships:**
    - An initial screening, using Pearson Correlation, reduced the dimensionality of explanatory variables from 244 to 125 variables.
    - Features related to real activity (industrial production, employment, consumption) display a strong positive association with GDP. In contrast, features related with categories such as interest rates, government finance and credit revealed more complex and nonlinear dynamics.
    - Analysing rolling correlation reveals that some relationships remain relatively stable over time, for example features related with employment and consumption, while others vary during recession periods.
- **PCA Insights:**
    - Principal Component Analysis (PCA) showed that a small subset of components captures most of the dataset's variance, with the first three components explaining over 80%, and seven components explaining more than 90%.
    - The first component, PC1, displayed a strong linear relationship with GDP, consistent in capturing the dominant variance direction.
    - The remaining higher-order components (PC2-PC8) have a weaker, more dispersed association with the target variable.

5 Data Preparation and Modelling

As described in the CRISP-DM methodology, the “Data Preparation” and the “Modelling” phases are closely related, often requiring refinements to the data preparation process so that the modelling techniques can be correctly applied. In the CRISP-DM introductory paper, the “Data Preparation” phase is described as covering “all activities to construct the final dataset (data that will be fed into the modelling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order” (Chapman *et al.*, 2000). In contrast, the same paper describes the “Modelling” phase as the stage in which various modelling techniques are selected and applied.

In the context of this dissertation, these two phases are treated jointly within a single chapter, as the preparation of the features and the definition of the modelling pipeline are interdependent and interconnected. Accordingly, this chapter covers several key aspects of the methodological framework, including the preprocessing steps applied to the dataset, the construction of lagged explanatory matrices, the strategy adopted for training and testing, the selection of forecasting models, and the common forecasting procedure applied across methods.

5.1 Data Preparation

As described in (Dataset Overview chapter), the initial dataset employed in the scope of this project is the FRED-QD, which contains a broad set of U.S macroeconomic and financial indicators at a quarterly frequency. The target variable in the scope of this machine learning task is real Gross Domestic Product (GDPC1), which serves as the basis for forecasting GDP dynamics and identifying recessionary periods.

Before modelling, a series of preprocessing steps were performed to ensure that the dataset was consistent and suitable for forecasting. These include the **removal of variables with missing values** and the **standardization of the explanatory variables**, transforming each series to **zero mean and unit variance**.

Regarding the preparation phase, it is important to note that the objective was not to optimize each individual model in isolation, but to establish a consistent and comparable dataset that could be used uniformly across all forecasting approaches. While certain models may benefit from specific preprocessing steps (for example differencing variables for SARIMAX or leaving raw features for tree-based methods), introducing model-specific adjustments would compromise the comparability of results. For this reason, the preprocessing was restricted to steps that are broadly applicable to all models. **The only exception** to this principle concerned the PCA-based strategy, where variables containing missing values were not discarded, but instead were handled by applying *forward fill* and *backward fill* imputation, as described in subchapter 4.7.

5.2 Feature Selection Strategies

Given the high dimensionality of the FRED-QD dataset, different feature selection and dimensionality reduction strategies were considered and tested. The main objective was to evaluate whether reducing dimensionality could improve forecast accuracy or computational efficiency. **Three** different strategies were implemented: usage of **all available features** without prior selection, **correlation-based filtering** and **Principal Component Analysis (PCA)**.

5.2.1 All Features (No Selection)

As a baseline, all cleaned and standardized explanatory variables were retained without additional filtering or dimensionality reduction. This approach preserved the entire information in the dataset, and while it increased the risk of including noisy predictors, testing this method was important to evaluate whether explicit feature selection methods such as correlation filtering or PCA provided benefits over using the full dataset. The dataset used for training, validation, and testing using this method contained **182** feature variables, as well as the target variable (GDPC1). The dataset can be consulted in Appendix A, Table 11

5.2.2 Correlation-Based Filtering (Pearson Correlation)

In this strategy, explanatory variables were selected based on their linear correlation with the target variable, real GDP (GDPC1). Pearson correlation coefficients were computed, and only those indicators exhibiting meaningful correlations were retained ($|r| > 0{,}90$, $\rho < 0{,}05$).

In addition, any remaining variables that still contained missing values were excluded from the feature set. The resulting dataset comprised **102** variables and the target variable (GDPC1), and is listed in Appendix A, Table 12.

5.2.3 Principal Component Analysis (PCA)

In this strategy, a PCA analysis was applied to the initial dataset, resulting in the extraction of 8 principal components. The cumulative explained variance was higher than 0.90, indicating that these components captured more than 90% of the total variability present in the original dataset. The resulting dataset comprised of the 8 PCA components as feature variables and the target variable (GDPC1).

5.3 Forecasting Models

Following the preparation phase and the definition of the feature selection strategies, several forecasting models were implemented and tested under a common framework. The main objective was to compare the performance of multiple machine learning models in relation to the task of forecasting the U.S GDP.

Four distinct models were selected for this purpose:

- **Linear Regression**: serves as a simple linear benchmark against which more complex models could be compared.
- **SARIMAX**: represents a statistical model commonly used in econometrics, capable of integrating external explanatory variables for robust time-series forecasting.
- **Random Forest**: nonparametric ensemble method designed to capture nonlinearities among the explanatory variables.
- **XGBoost (Extreme Gradient Boosting):** boosting-based tree ensemble that demonstrates strong predictive performance in high-dimensional datasets, while also being computationally efficient during training.

The evidence reported in the existing literature and their methodological diversity informed the selection of these four forecasting models.

Linear Regression was included as a baseline benchmark. Several previous papers, (Nyman and Ormerod, 2017); (Sties, 2017); (Chung, 2023), have used regression-based models such as OLS, Ridge or Logistic Regression as starting points for comparison. These models provide interpretability and establish a reference against which more complex techniques can be evaluated.

SARIMAX represent a statistical time-series model that integrates autoregressive and moving average components with exogenous regressors. Models of ARIMA/VAR family have been widely used in past literature for macroeconomic forecasting (McCracken and Ng, 2021)**.**

Random Forest was selected as a representative ensemble learning method. The literature shows strong evidence of its effectiveness in economic forecasting. For example, (Nyman and Ormerod, 2017, 2020) demonstrate that Random Forest outperforms OLS for multi-step forecasts in the US and UK. (Zyatkov and Krivorotko, 2021b) also highlighted its competitive performance relative to logistic regression and boosting models. Random Forest seems particularly effective at capturing nonlinearities and interactions in high-dimensional structure of the dataset used in the context of this project.

XGBoost (Extreme Gradient Boosting) was chosen as an advanced boosting method, combining strong predictive performance in high-dimensional datasets, while also being computationally efficient. The literature consistently highlights its accuracy in economic forecasting. For instance, (Qilu, 2022) reports that XGBoost achieves AUROC values above 0.99 when forecasting recessions across advanced economies, outperforming traditional logit benchmarks.

5.4 Forecasting Methodology

The aim of this subchapter is to clarify the reader of the methodological framework adopted to generate the forecasts that will be detailed in the next chapter 6, “Results and Discussi”. The main objective is to establish a consistent methodology that could be applied across different models and feature selection strategies, ensuring comparability of results.

5.4.1 Forecasting Strategy

The forecasting task was formulated as a **multi-step time-series prediction problem**, in which the target variable was the real GDP (GDPC1). To address this problem, two main forecasting strategies are widely discussed in the literature:

- The **direct strategy**, which trains a separate model for each forecast horizon.
- The **recursive strategy**, which generates longer-horizon forecasts by repeatedly feeding back the model’s own predictions.

In the scope of this project, a **recursive multi-step strategy** was adopted. The predictor variables present in the dataset were incorporated as exogenous variables, with the assumption

that their values are known at the time each forecast is generated. At each forecast origin, the model is trained using all available data up to that point, and forecasts are generated step by step by feeding the predicted values back into the model.

For example, to forecast three steps ahead, the model first predicts y_{t+1} using past lags and the exogenous variables x_{t+1} (assumed known), then uses this prediction together with x_{t+2} to obtain y_{t+2}, and so on until y_{t+3} is forecasted. In Figure 22, we can observe an example of including exogenous variables as predictors in the forecasting process.

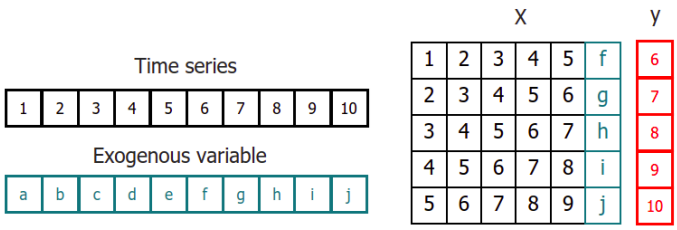

Figure 22 - Time Series Forecasting including exogenous variables. Source: (Rodrigo and Ortiz, 2024b)

This methodology reflects real-time forecasting practice in macroeconomics and has been previously used in the literature, for example in (Nyman and Ormerod, 2017).

5.4.2 Data Partitioning and Recession-Oriented Test Window

To preserve the temporal structure of the data and avoid look-ahead bias, observations were split chronologically into **training**, **validation**, and **testing** sets. Unlike conventional random or proportional splits, validation and testing windows were aligned with specific calendar periods coinciding with major U.S. recessions as defined by the NBER. In each case, the **validation** set corresponds to the quarters immediately preceding the **test** window and was used exclusively for model selection and hyperparameter tuning; the **test** set was kept fully unseen for final evaluation. This setup was chosen to align with the project's main objective: assessing model effectiveness specifically during recession periods.

- **Period 1** – 1973-1975 Recession (Oil Crisis):
  - **Train**: 1959Q1 – 1970Q4
  - **Validation**: 1971Q1 – 1973Q3
  - **Test**: 1973Q4–1976Q4
- **Period 2** - Early 1980s double-dip recessions:
  - **Train**: 1959Q1–1976Q4
  - **Validation**: 1977Q1–1980Q3
  - **Test:** 1980Q4–1983Q4
- **Period 3** - 2008 Great Recession:
  - **Train**: 1959Q1–2003Q4
  - **Validation**: 2004Q1–2007Q3

- **Test:** 2007Q4–2010Q4

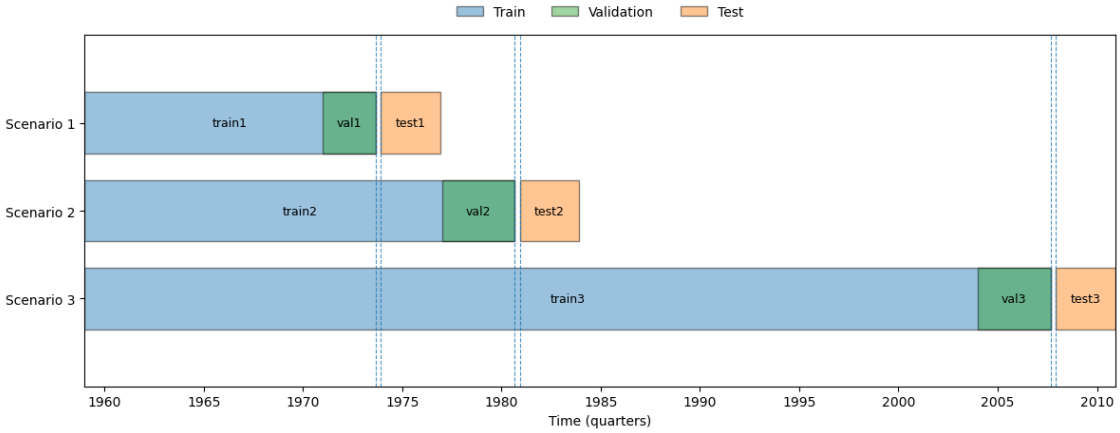

Figure 23 - Timeline representation of the training, testing and validation periods selected for the forecasting task.

5.4.3 Validation Strategy and Hyperparameter Optimization

To prevent look-ahead bias and keep evaluation faithful to real-time forecasting, hyperparameter tuning was performed only on the validation window immediately preceding each test window. Rolling-origin cross validation was used on the train plus validation segment with an expanding window. At each fold, the model was refit on all data up to the fold origin and evaluated using Mean Absolute Error as the selection metric.

The cross-validation scheme used **3 folds**, meaning that each model–parameter–lag configuration was evaluated across three consecutive forecast origins inside the validation window, and the average MAE determined the best configuration.

To implement this validation strategy, **skforecast**'s `grid_search_forecaster` was used for the ML models (Linear Regression, Random Forest, XGBoost) and `grid_search_sarimax` for the statistical model (SARIMAX). The implementation is provided in Code Snippet 2

```
results_grid = grid_search_forecaster(
    forecaster=forecaster,
    y=y_trval,
    lags_grid=lags_list,
    param_grid=param_grid,
    steps=steps_per_fold,
    refit=True,
    metric='mean_absolute_error',
    initial_train_size=initial_train_size,
    exog=X_trval,
    fixed_train_size=False,
    allow_incomplete_fold=True,
    return_best=True,
    verbose=False
```

)

Code Snippet 2 - Validation Strategy using grid_search_forecaster

Within the described validation strategy, multiple hyperparameter combinations were evaluated. These are summarized in Table 6.

Table 6 - Hyperparameter grids and lag sets used for validation-only tuning (rolling-origin CV, 3 folds; selection metric = MAE)

Model	Tuned Hyperparameters	Values Tested	Lags	Space Size (Nº different models tested)
SARIMAX	order	[(0, 1, 0), (1, 1, 0), (1, 1, 1), (2, 1, 0), (2, 1, 1), (2, 2, 1), (3, 1, 0), (3, 1, 1), (3, 2, 1),]	----	9
	seasonal_order	(0, 0, 0, 0)		
Linear Regression	fit_intercept	[True, False]	[2,4,6,12]	16
	positive	[True, False]		
Random Forest	n_estimators	[50,100,200]	[2,4,6,12]	36
	max_depth	[5,10,15]		
XGBoost	n_estimators	[200,400]	[2,4,6,12]	256
	max_depth	[2,4]		
	learning_rate	[0.10, 0.05]		
	min_child_weight	[5,10]		
	subsample	[0.5, 0.9]		
	colsample_bytree	[0.3, 0.7]		

5.4.4 Implementation Details

The implementation of the forecasting pipeline was carried out in **Python**, chosen for its large ecosystem of data science libraries. The data manipulation and preprocessing relied mainly on two libraries: **pandas** and **NumPy,** while visualization tasks used **matplotlib.** As for the forecasting models, these were implemented using well known libraries, such as **statsmodels** (SARIMAX), **scikit-learn** (Linear Regression and Random Forest), and **xgboost** (Gradient Boosting).

In order to facilitate the implementation of the recursive multi-step strategy with the machine learning models, **skforecast** was used. This library provides utilities for generating lagged features, which are necessary to integrate exogenous predictors into the forecasting process. Table 7 provides an overview of the python libraries used in the implementation phase of the forecasting pipeline previously described.

Table 7 – Summary of Python libraries used in the implementation of forecasting pipeline

Library	Purpose
pandas	Data manipulation, handling time series indices, and overall preprocessing tasks
NumPy	Numerical operations
matplotlib	Visualization of time series and forecast results
statsmodels	Implementation of SARIMAX
scikit-learn	Implementation of Linear Regression and Random Forest; evaluation metrics (RMSE, MAE)
xgboost	Gradient boosting model
skforecast	Utilities for recursive forecasting with ML models, including lagged feature generation and recursive multi-step predictions

To ensure methodological consistency and reproducibility, four different python functions were developed to represent the forecasting task for each model: SARIMAX, Linear Regression, Random Forest and XGBoost. Each function consists of the complete workflow: data preparation, model fitting, recursive forecasting, performance evaluation and visualization. The functions follow the same underlying structure, only differencing in the specific model used in the fitting and prediction steps.

6 Results and Discussion

This chapter presents the results obtained from the forecasting experiments carried out in the scope of this project and detailed in the previous chapter. This chapter is organized into four main parts. First, the results are reported separately for each test period, in line with the objective assessing each model's performance during specific phases of economic recessions. Next, a cross-period comparison is conducted, to provide a global overview of the performance across the three time periods. This is followed by the "Discussion" subchapter, where the results are interpreted considering existing literature and methodological considerations. Finally, the chapter concludes with a reflection on the main limitations and practical considerations in the scope of the findings.

6.1 Results

In the scope of this project, the forecasting task was defined as a regression problem, where the primary objective is to forecast the numeric value of the target variable (GDPC1). The evaluation is carried out using regression metrics, namely the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE).

The **Regression Metrics** are directly calculated by comparing the forecasts with the actual values, and consist of the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE):

- **Root Mean Squared Error (RMSE)**: measures the square root of the average squared difference between predicted values ($\hat{y}_t$) and actual values (y_t):

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

- **Mean Absolute Error (MAE)**: measures the average absolute difference between predicted and actual values:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

MAE represents the average magnitude of errors. RMSE gives a higher penalty on larger errors and generally reflects greater discrepancies in predictions than MAE. Both are in the same units as the original values, giving a straightforward interpretation.

6.1.1 Results by Time Periods

This subchapter presents GDP-forecasting results for the three test windows—1973Q4–1976Q4, 1980Q4–1983Q4, and 2007Q4–2010Q4. For each period, **the best-performing configuration** for every (Model × Feature Strategy) pair is reported, selected via the rolling-origin validation methodology, including hyperparameter tuning, described in 5.4.3.

Test-window performance is summarized with **RMSE** and **MAE**, calculated between the predicted values and the real GDPC1. Compute Time (s) reflects the training, validation, and testing time for the chosen configuration, providing a concise view of computational cost.

6.1.1.1 Period 1: 1973-1976 (Oil Crisis)

Table 8 - Results Period 1 - 1973-1976 (Oil Crisis)

Model	Feature Strategy	Best Model Configuration	RMSE	MAE	Compute Time(s)
SARIMAX	**All Features**	order=(1,1,0), seasonal_order=(0,0,0,0)	59.13	49.17	103.65
	Corr. Filter (Pearson)	order=(2,1,0), seasonal_order=(0,0,0,0)	**38.56**	**35.68**	24.73
	PCA	order=(3,2,1), seasonal_order=(0,0,0,0)	147.94	135.03	15.28
Linear Regression	**All Features**	Lags=6 fit_intercept=True positive=True	99.89	93.65	0.38
	Corr. Filter (Pearson)	Lags=4 fit_intercept=True positive=False	**11.45**	**9.05**	0.31
	PCA	Lags=4 fit_intercept=True positive=False	248.32	222.84	0.24
Random Forest	**All Features**	Lags=4 max_depth=5 n_estimators=50	594.61	568.06	65.68
	Corr. Filter (Pearson)	Lags=4 max_depth=5 n_estimators=50	**238.97**	**196.58**	41.6
	PCA	Lags=6 max_depth=15 n_estimators=100	544.83	502.44	12.85
XGBoost	**All Features**	Lags=6 colsample_bytree=0.7 learning_rate=0.1 max_depth=4 min_child_weight=5 n_estimators=400 subsample=0.9	223.23	159.15	187.90
	Corr. Filter (Pearson)		**195.81**	**146.49**	144.45

		Lags=2 colsample_bytree=0.3 learning_rate=0.1 max_depth=4 min_child_weight=5 n_estimators=400 subsample=0.9			
	PCA	Lags=6 colsample_bytree=0.7 learning_rate=0.1 max_depth=4 min_child_weight=5 n_estimators=400 subsample=0.9	245.54	180.60	70.03

As can be observed in Table 8, the highest scoring configuration for each model, is highlighted in bold. The results for the test period between 1973Q4 and 1976Q4, which encompasses the Oil Crisis recession, reveal clear differences across model configurations and feature selection strategies. Overall, the best Linear Regression and SARIMAX model configurations, using predictors filtered through Pearson Correlation, achieved the lowest RMSE and MAE.

The best Linear Regression model configuration, when using predictors filtered through Pearson correlation and 6 lags, recorded RMSE and MAE values of 11.45 and 9.05. The SARIMAX model also demonstrated solid performance when incorporating predictors filtered through Pearson correlation, achieving RMSE and MAE values of 38.56 and 35.68, respectively.

Figure 24 and Figure 25 provide a visual representation of the forecasting results for the best Linear Regression and SARIMAX models respectively.

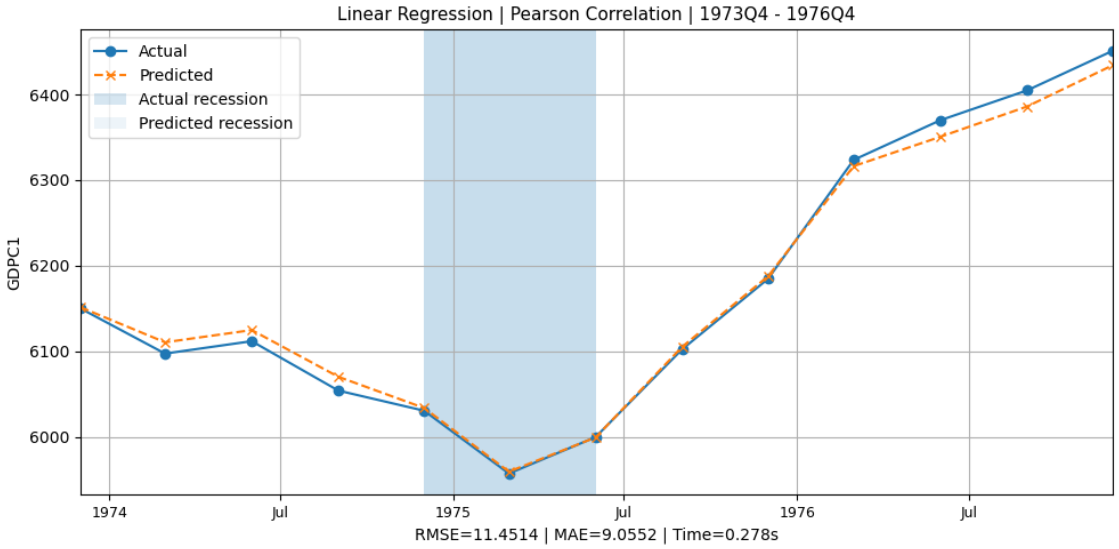

Figure 24 – Results Linear Regression for the test period 1973Q4–1976Q4, based on correlation-filtered predictor variables. Shaded area indicates the actual recession period (Oil Crisis).

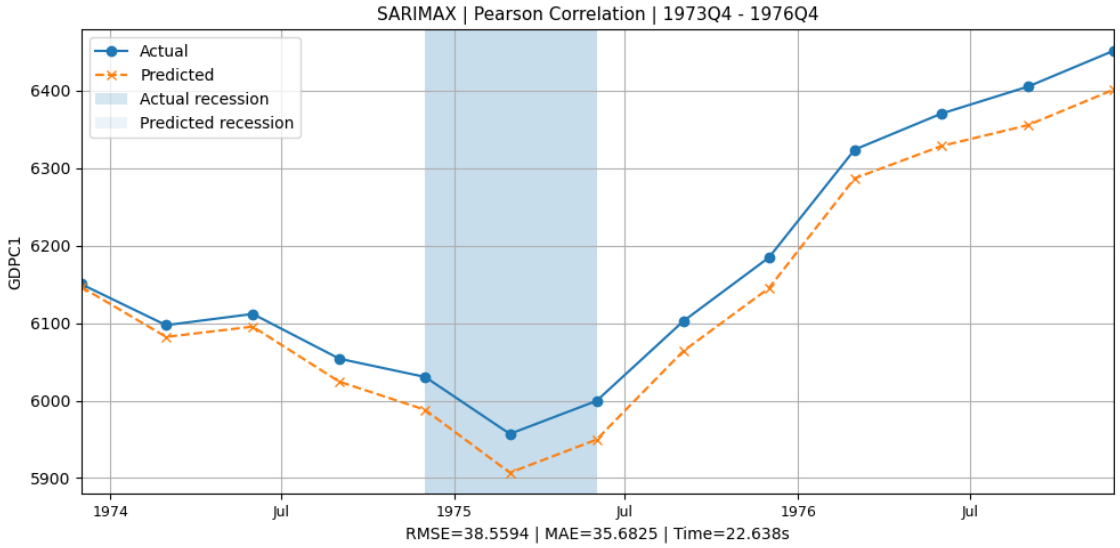

Figure 25 - Results SARIMAX the test period 1973Q4–1976Q4, based on correlation-filtered predictor variables. Shaded areas indicate the actual recession period (Oil Crisis).

As for feature importance, we can observe in Figure 26 and Figure 27 below, the most important features in the forecasting task for the best Linear Regression and SARIMAX model configurations, respectively. For Linear Regression, the variable OPHPBS (Real Output Per Hour of All Persons) revealed to be the most important, while for SARIMAX, variable GPDCI1 (Real Gross Private Domestic Investment) was the most relevant.

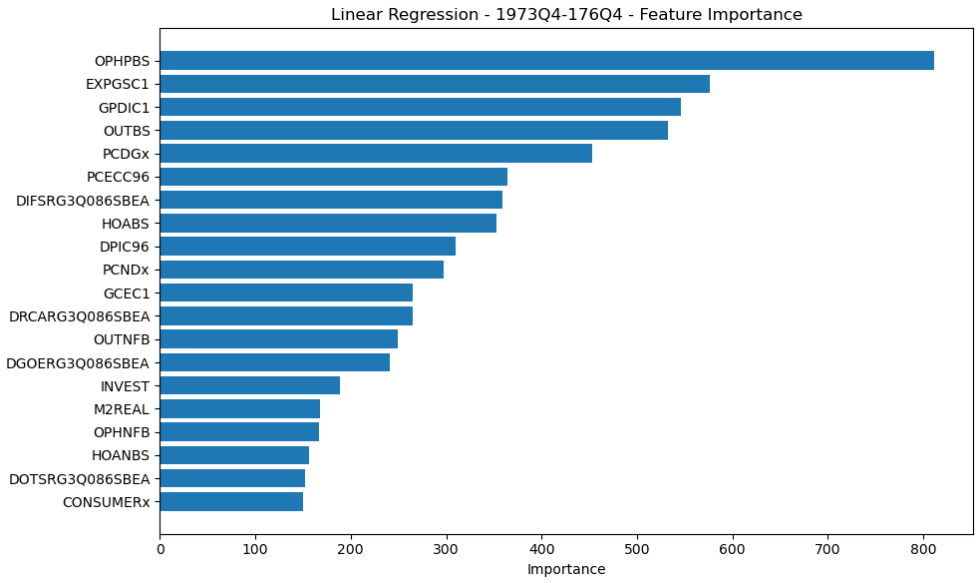

Figure 26 - Linear Regression Feature Importance - 1973Q4-1976Q4

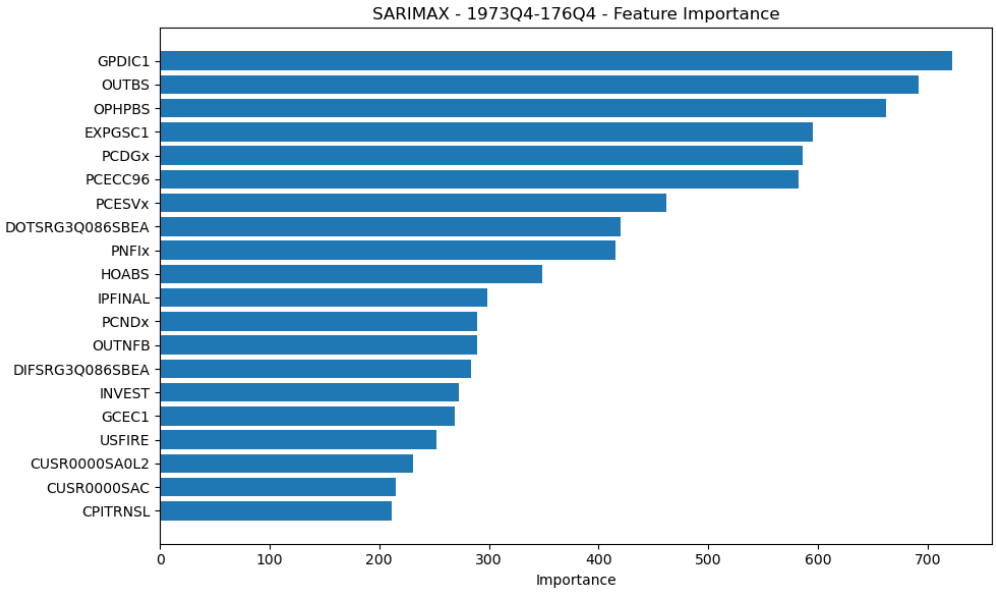

Figure 27 - SARIMAX Feature Importance - 1973Q4-1976Q4

6.1.1.2 Period 2: 1980-1983 (Double-dip recession)

Table 9 - Results Period 2 - 1980-1983 (Double-dip recession)

Model	Feature Strategy	Best Model Configuration	RMSE	MAE	Compute Time(s)
SARIMAX	**All Features**	order=(0,1,0), seasonal_order=(0,0,0,0)	52.70	46.35	113.94
	Corr. Filter (Pearson)	order=(1,1,0), seasonal_order=(0,0,0,0)	**29.18**	**23.60**	25.07
	PCA	order=(1,1,0), seasonal_order=(0,0,0,0)	62.47	52.70	20.21
Linear Regression	**All Features**	Lags=2 fit_intercept=True positive=True	28.23	23.89	0.58
	Corr. Filter (Pearson)	Lags=2 fit_intercept=True positive=True	**13.83**	**10.80**	0.29
	PCA	Lags=12 fit_intercept=True positive=False	61.94	54.34	0.24
Random Forest	**All Features**	Lags=2 max_depth=10 n_estimators=50	400.31	356.21	116.82
	Corr. Filter (Pearson)	Lags=2 max_depth=10 n_estimators=50	**291.43**	**239.57**	63.45
	PCA	Lags=6 max_depth=10 n_estimators=50	360.69	289.86	16.34
XGBoost	**All Features**	Lags=6 colsample_bytree=0.7 learning_rate=0.1 max_depth=4 min_child_weight=5 n_estimators=400 subsample=0.9	264.31	215.82	229.47
	Corr. Filter (Pearson)	Lags=4 colsample_bytree=0.3 learning_rate=0.1 max_depth=4 min_child_weight=5 n_estimators=400 subsample=0.9	**257.38**	**206.14**	175.30
	PCA	Lags=6 colsample_bytree=0.7 learning_rate=0.1 max_depth=4 min_child_weight=5 n_estimators=400 subsample=0.9	269.05	198.21	75.56

As can be observed in Table 9, the highest scoring configuration for each model, based on both error metrics, is highlighted in bold. The results for the testing window between 1980Q4 and 1983Q4, which includes the double-dip recession of the early 80s as well as the period of economic recovery following that, reveal that, once again, both Linear Regression and SARIMAX stand out with the best results. The best Linear Regression model configuration, using Pearson based correlation filtering feature strategy, achieved the lowest RMSE (13.83) and the lowest MAE (10.80). SARMAX also displayed good performance, most notably when using correlation-based feature selection, having an RMSE and MAE of 29.19 and 23.60, respectively. On the other hand, Random Forest and XGBoost models provided weaker results, with higher error values.

As in the previous subchapter, for readability purposes, only the visualization of the results of the best-performing models are presented in following figures: Figure 28 and Figure 29

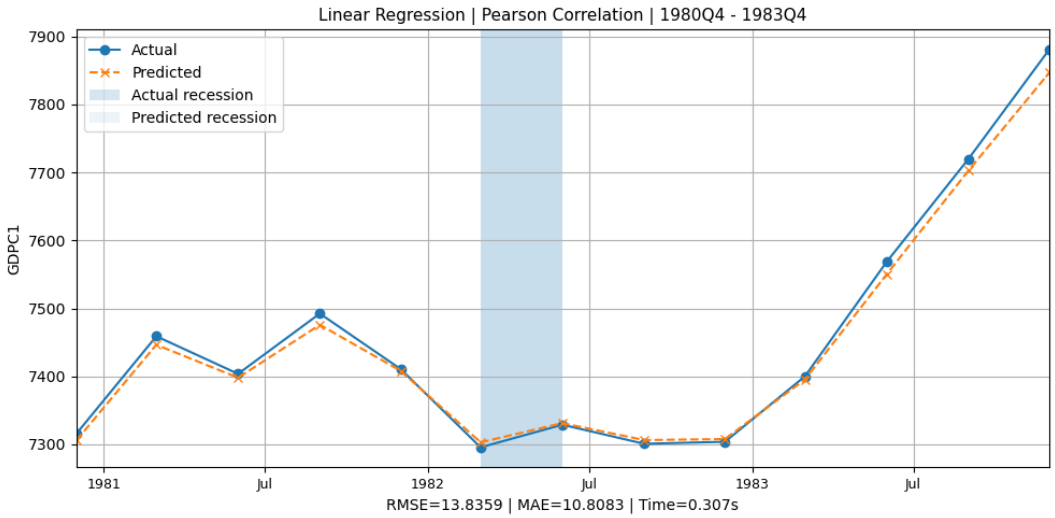

Figure 28 - Results Linear Regression for the test period 1980Q4–1983Q4, based on correlation-filtered predictor variables. Shaded area indicates the actual recession period.

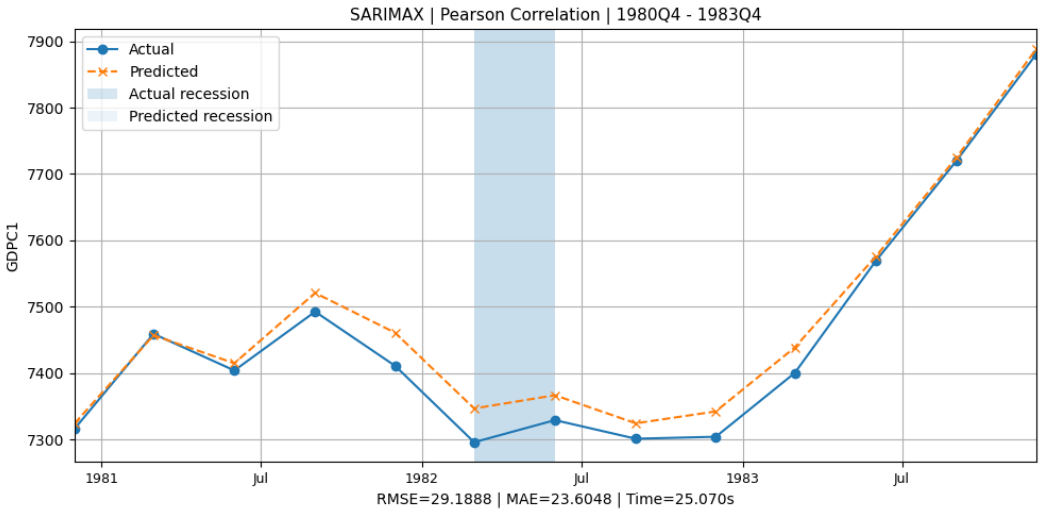

Figure 29 - Results SARIMAX for the test period 1980Q4–1983Q4, based on correlation-filtered predictor variables. Shaded area indicates the actual recession period.

As for feature importance, we can observe in Figure 30 and Figure 31 below, the most important features for Linear Regression and SARIMAX model configurations in the 1980Q4-1983Q4 testing period, respectively. For SARIMAX, the variable PCECC6 (Real Personal Consumption Expenditures (PCE) revealed to be the most important, while for Linear Regression, variable OUTBS (Real Value-Added Output for All Workers) was the most relevant.

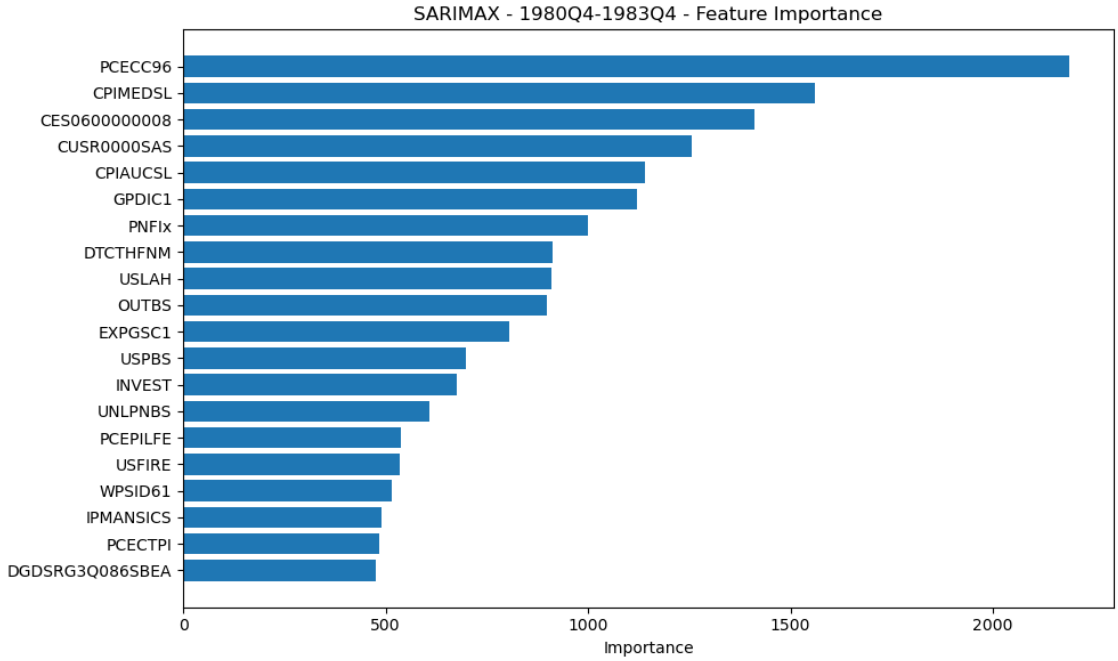

Figure 30 - SARIMAX Feature Importance - 1980Q4-1983Q4

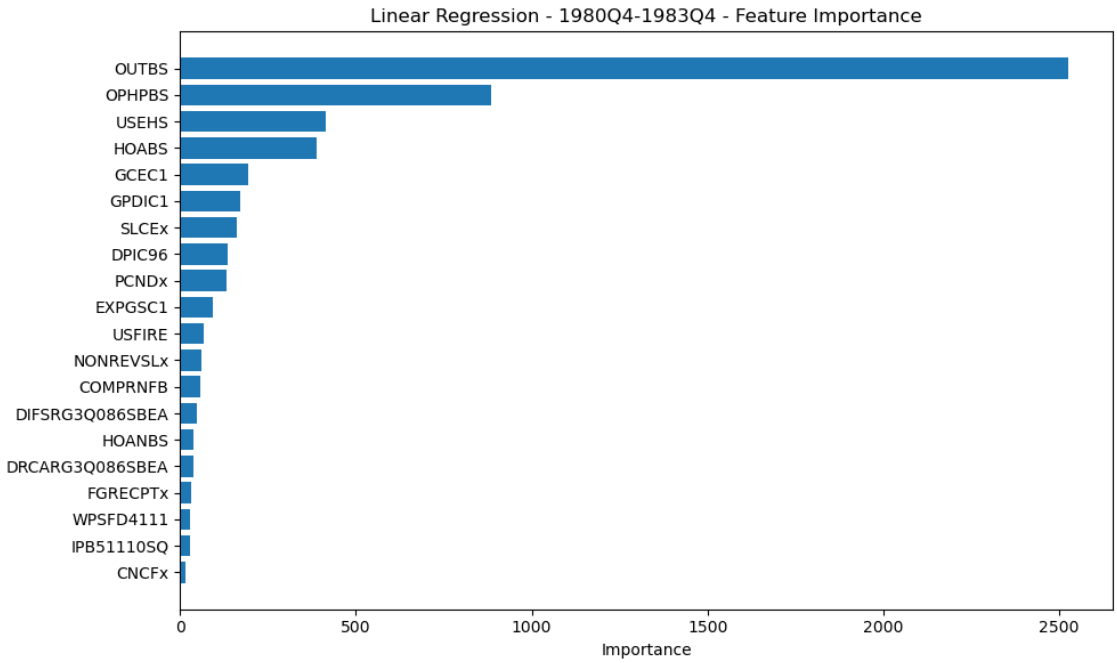

Figure 31 - Linear Regression Feature Importance - 1980Q4-1983Q4

6.1.1.3 Period 3: 2007-2010 (Great Recession)

Table 10 - Results Period 3 - 2007-2010 (Great Recession)

Model	Feature Strategy	Best Model Configuration	RMSE	MAE	Compute Time(s)
SARIMAX	**All Features**	order=(1,1,1), seasonal_order=(0,0,0,0)	699.50	598.39	233.04
	Corr. Filter (Pearson)	order=(1,1,0), seasonal_order=(0,0,0,0)	**74.08**	**67.69**	215.31
	PCA	order=(2,1,0), seasonal_order=(0,0,0,0)	130.89	100.18	32.06
Linear Regression	**All Features**	Lags=2 fit_intercept=True positive=True	104.27	97.09	0.45
	Corr. Filter (Pearson)	Lags=12 fit_intercept=True positive=False	**79.10**	**68.33**	0.50
	PCA	Lags=12 fit_intercept=True positive=True	991.66	858.49	0.22
Random Forest	**All Features**	Lags=4 max_depth=15 n_estimators=100	1321.87	1180.91	320.57
	Corr. Filter (Pearson)	Lags=4 max_depth=10 n_estimators=100	636.48	569.96	175.62
	PCA		**551.14**	**486.28**	33.19

		Lags=12 max_depth=15 n_estimators=50			
XGBoost	**All Features**	Lags=12 colsample_bytree=0.3 learning_rate=0.1 max_depth=4 min_child_weight=5 n_estimators=400 subsample=0.9	228.70	209.38	387.98
	Corr. Filter (Pearson)	Lags=2 colsample_bytree=0.7 learning_rate=0.1 max_depth=4 min_child_weight=5 n_estimators=400 subsample=0.9	237.17	201.20	195.56
	PCA	Lags=12 colsample_bytree=0.3 learning_rate=0.1 max_depth=4 min_child_weight=5 n_estimators=400 subsample=0.9	**229.81**	**193.46**	88.03

As can be observed in Table 10, the highest-scoring configuration for each model, based error metrics, is highlighted in bold. The results for the testing window between 2007Q4 and 2010Q4, which encompasses the Great Recession and the subsequent early recovery, indicate that both Linear Regression and SARIMAX once again achieved the strongest outcomes. With correlation-based feature selection, the SARIMAX model delivered the lowest RMSE (74.08) and MAE (67.69). Linear Regression also performed well, particularly when using correlation-based predictors, achieving an RMSE of 79.10 and an MAE of 68.33. By contrast, Random Forest and XGBoost models generally produced higher error levels.

As in the previous subchapter, for readability, only the plots of the best-performing models are presented in the following figures (Figure 32 and Figure 33).

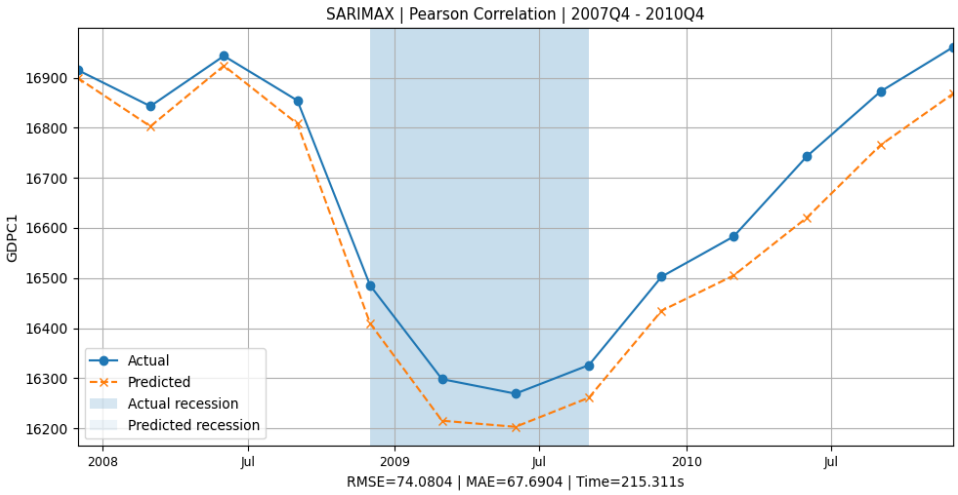

Figure 32 - Results SARIMAX for the test period 2007Q4–2010Q4, based on correlation-filtered predictor variables. Shaded area indicates the actual recession period.

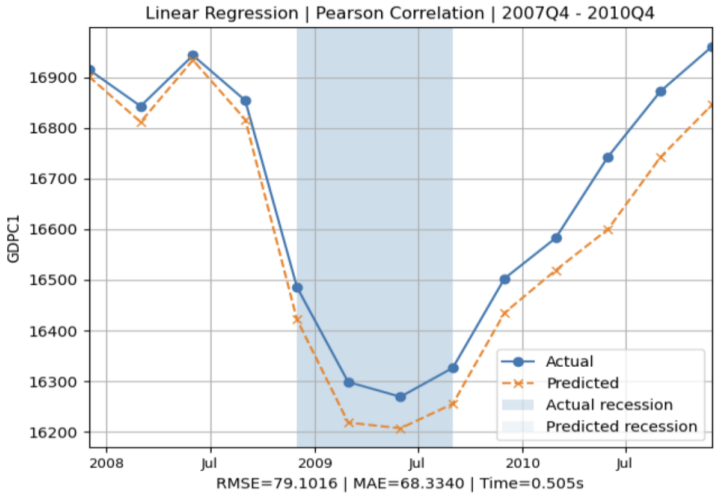

Figure 33 - Results Linear Regression for the test period 2007–2010Q4, based on correlation-filtered predictor variables. Shaded area indicates the actual recession period.

In relation to feature importance, we can observe in Figure 34 and Figure 35 below, the most important features for SARIMAX and Linear Regression models in the 2007Q4-2010Q4 testing period, respectively. Analysing both figures, we can observe that for SARIMAX, PAYEMS (All Employees: Total Nonfarm Payrolls) was the most relevant feature. While for Linear Regression, USPRIV (All Employees: Total Private Industries) was the most relevant feature.

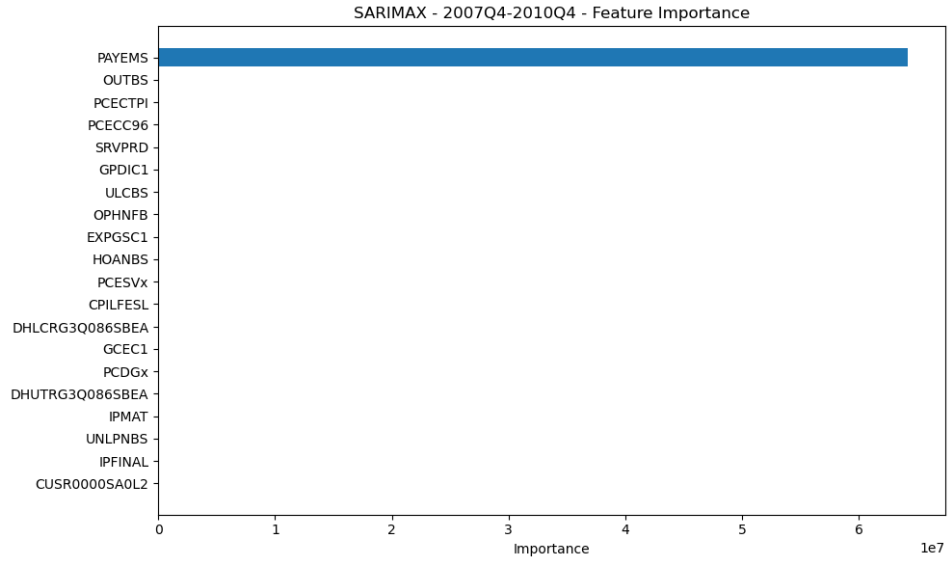

Figure 34 - SARIMAX Feature Importance - 2007Q4-2010Q4

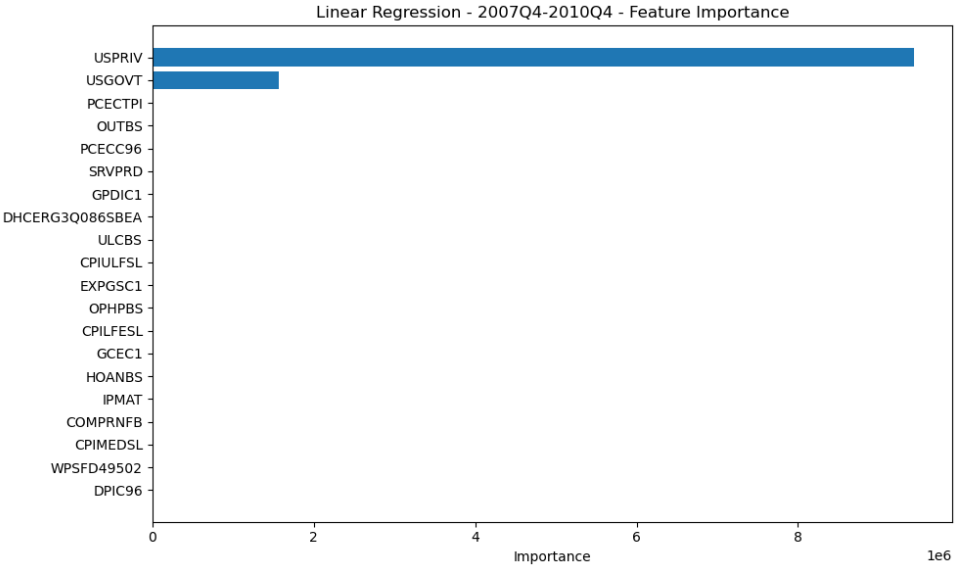

Figure 35 – Linear Regression Feature Importance - 2007Q4-2010Q4

6.2 Discussion

The comparative results reveal some clear insights into the performance of different types of statistical and machine learning models across the three periods considered in this study. Overall, a consistent pattern emerges when analyzing the results: Linear Regression and SARIMAX outperform nonlinear models, such as Random Forest and XGBoost, in terms of RMSE and MAE. Both Linear Regression and SARIMAX seem to be particularly sensitive to different feature selection strategies, especially benefiting from the Pearson Correlation Filtering strategy. For instance, in the 1980Q4–1983Q4 test period, the RMSE of Linear Regression dropped from 28.23 when using all features to 13.83, when trained on variables filtered by correlation. This improvement reflects the reduction of noise achieved by excluding weakly correlated predictors, which enhances forecasting accuracy. By contrast, PCA feature strategy seems to consistently worsen the results throughout the three testing periods, this can perhaps be explained by the fact that the principal components aim at maximizing overall variance and therefore can dilute underlying economic signals.

In clear contrast to Linear Regression and SARIMAX, Random Forest and XGBoost, generally underperformed across the three testing periods. Random Forest yielded consistently high error levels. This can be somewhat explained by the limited sample size of quarterly macroeconomic data, the high dimensionality of the predictor set, and the presence of strong autocorrelation structures that tree-based methods do not explicitly capture. Applying feature selection strategies on these models helped mitigate these issues to some extent and, overall, improving the forecasting results. For example, during the 1973-1976 period, Random Forest achieved an RMSE of 238.97, when using Pearson Correlation Filtering, a substantial improvement over the 594.61 obtained when using the full set of predictors.

When it comes to feature importance, the results presented provide evidence that the predictive relevance of macroeconomic indicators seems to fluctuate between recession periods and are not consistent across time but instead fluctuate depending on the recession period considered. For example, in the 1973Q4-1976Q4 window, both Linear Regression and SARIMAX assigned high weights to variables associated with categories “Earnings and Productivity” and “Investment” (OPHPBS and GPDCI1). On the other hand, in the 2007Q4-2010Q4 testing window, that covers the Great Recession, SARIMAX gives a stronger emphasis on variables related with the job market (PAYEMS).

6.2.1 Limitations and Considerations

A few limitations should be considered when interpreting the results of this study. First, the dataset used consists of relatively few observations at a quarterly frequency, while at the same time being highly dimensional, which increases the risk of overfitting and makes train/test splits more challenging. This limitation is inherent to the context of the problem addressed in this dissertation, since macroeconomic variables are typically published at a monthly or, as in this

case, quarterly frequency. Even when extending the training data back to 1959, the number of available observations for each train/test split remains relatively low.

Second, there is an inherent imbalance in the target variable. Given the nature of recessions, periods of non-recession, that is, of economic expansion, are significantly more frequent when compared to periods of economic recession.

Third, the forecasting design assumes that all the macroeconomic indicators are known at the time of prediction. In practice, many of these indicators are published with a lag, often at the same time as the GDP target variable and are subject to subsequent statistical revisions. This creates a discrepancy between the experimental setup and the information set that would be available in real-time forecasting. Furthermore, the evaluation was limited to three test windows, which, although chosen to represent different recessionary contexts, may not fully capture the robustness of the models in other macroeconomic environments.

Finally, the hyperparameter optimization procedures, while extensive, cannot guarantee that the models reached the optimal result, particularly for complex methods with multiple hyperparameters such as Random Forest and XGBoost.

These considerations should be kept in mind when interpreting the results and when assessing the transferability of the findings to real-time forecasting applications.

7 Conclusion

This chapter summarizes the main achievements and contributions of the study, reflects on the challenges encountered during the project's development, and concludes by outlining directions for future research along with final considerations.

7.1 Achievements and Contributions

As stated in the introductory chapter, the core research objective of this project was to develop a machine learning model capable of effectively predicting the onset of economic recessions.

One of the main contributions of this dissertation lies in the comparative assessment of different modelling approaches. Both statistical models (SARIMAX) and machine learning models (Linear Regression) were implemented and evaluated across three historical recession periods (1973–1976, 1980–1983, and 2007–2008). This comparative approach allowed for a comprehensive understanding of the relative strengths and weaknesses of each approach in the scope of this project.

Another significant achievement concerns the implementation and testing of different feature selection strategies, namely Pearson Correlation Filtering and Principal Component Analysis (PCA). The results showed that the choosing of specific feature selection strategies can have a direct impact on model performance, with Pearson Correlation Filtering consistently improving predictive accuracy across the different testing periods.

Another contribution to the field was the study of the feature relevance throughout different recession periods. The analysis of the feature importance on the best models in each period proves that the predictive relevance of macroeconomic indicators is not consistent throughout the different periods but rather changes and evolves depending on the specific recessionary context.

Overall, both SARIMAX and Linear Regression proved to be effective approaches to forecast periods of economic recession, with relatively low RMSE and MAE values in the testing periods considered. These findings directly align with the core research objective of this dissertation: to evaluate and compare the effectiveness of machine learning models in predicting the onset of economic recessions

7.2 Challenges and limitations

Throughout the development of the work, there were several challenges and limitations.

One of the main challenges was related to the dataset characteristics, as it consisted of a relatively small sample with high dimensionality. This limitation stems from the inherent nature of the project's scope, since macroeconomic variables are typically reported at low frequencies, such as monthly or quarterly, which restricts the number of available observations. Furthermore, the dataset also presented a natural imbalance in the target variable, with significantly fewer recession periods compared to non-recession periods. These characteristics significantly increased the complexity of the modelling approach, constraining the design of testing strategies.

A further difficulty concerned the framing of the forecasting problem itself. Recession prediction can be approached either as a classification task, predicting whether a given period will be recessionary or not, or as a regression task, where the objective is to forecast GDP growth and infer recession periods from negative growth patterns. Much of the existing literature and methodological support in time series forecasting is oriented toward regression tasks, this project ultimately adopted the regression-based approach as the most consistent and feasible strategy for addressing the research problem.

Another challenge was the limited background knowledge in the field of economics. Although the author has a personal interest in following economic developments, the lack of formal training in economics made it difficult to extract deeper domain-specific insights from the results. As this dissertation was primarily developed within the scope of data science, the focus remained on the methodological and technical aspects of forecasting, rather than on deriving extensive economic interpretations of the outcomes.

7.3 Future work

Given the challenges and limitations faced, future work can focus on addressing the issue of a relatively small sample size and the inherent imbalance recession data. One possible strategy to address this can be to explore time series data augmentation techniques. As described by (Wen *et al.*, 2022), such methods can include decomposition approaches (e.g., trend and seasonal decomposition), embedding space transformations, deep generative models, and automated data augmentation pipelines.

Additionally, future work could benefit from the involvement of economic researchers who can review the findings and provide deeper domain-specific interpretations. This collaboration would help bridge the gap between data science and economics, ensuring that the methodological results are translated into insights that are meaningful and useful for the broader economic community.

7.4 Final considerations

This dissertation represents a significant milestone in the academic journey of the author. The theme was chosen as it bridges two areas of personal interest. On one hand, economics has always been a field that sparked curiosity, particularly in understanding how macroeconomic dynamics influence societies. On the other hand, the field of data science provided the methodological and analytical tools necessary to address complex forecasting challenges.

In conclusion, this dissertation provided crucial insights into how to conduct a rigorous methodological approach within the scope of time series forecasting. Throughout the development of this project, some advancements were introduced in the field, particularly in the systematic comparison of machine learning models for recession prediction, the evaluation of different feature selection strategies, and the analysis of feature relevance across distinct recessionary periods.

References

Aguiar-Conraria, L., Bação, P., Horta Correia, I., Alberto Ferreira, J., Reis, R., Tavares, J., Valério, N., *et al.* (2023), *Crises Na Economia Portuguesa : De 1910 a 2022*, Fundação Francisco Manuel dos Santos, Lisboa.

Alharbi, F.R. and Csala, D. (2022), "A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach", *Inventions*, Vol. 7 No. 4, p. 94, doi: 10.3390/inventions7040094.

Bell, F. and Smyl, S. (2018), "Forecasting at Uber: An Introduction", 6 September, available at: https://www.uber.com/en-PT/blog/forecasting-introduction/ (accessed 28 September 2025).

Board of Governors of the Federal Reserve System (US). (2025a), "Nominal Advanced Foreign Economies U.S. Dollar Index", Federal Reserve Bank of St. Louis.

Board of Governors of the Federal Reserve System (US). (2025b), "U.S. Dollars to Euro Spot Exchange Rate", Federal Reserve Bank of St. Louis.

Breiman, L. (2001), "Random Forests", *Machine Learning*, Vol. 45 No. 1, pp. 5–32, doi: 10.1023/A:1010933404324.

Brian Bell. (2015), "Do recessions increase crime?", 4 March, available at: https://www.weforum.org/stories/2015/03/do-recessions-increase-crime/ (accessed 29 December 2024).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000), *CRISP-DM 1.0: Step-by-Step Data Mining Guide*.

Chatfield, C. and Xing, H. (2019), *The Analysis of Time Series*, Chapman and Hall/CRC, Seventh edition. | Boca Raton, Florida : CRC Press, [2019] |, doi: 10.1201/9781351259446.

Chung, S. (2023), "Inside the black box: Neural network-based real-time prediction of US recessions".

Claessens, S. and KOSE, M.A. (n.d.). "Recession: When Bad Times Prevail", *Back to Basics*, available at: https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/Recession (accessed 29 December 2024).

"CÓDIGO DE CONDUTA — P.PORTO | Ensino Superior Público". (n.d.). , available at: https://www.ipp.pt/sobre/transparencia-integridade-anticorrupcao/codigo-de-conduta (accessed 29 December 2024).

Dickey, D.A. and Fuller, W.A. (1979), “Distribution of the Estimators for Autoregressive Time Series With a Unit Root”, *Journal of the American Statistical Association*, JSTOR, Vol. 74 No. 366, p. 427, doi: 10.2307/2286348.

“Ethics guidelines for trustworthy AI | Shaping Europe’s digital future”. (n.d.). , available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed 29 December 2024).

Guerra, O. and Eboreime, E. (2021), “The impact of economic recessions on depression, anxiety, and trauma-related disorders and illness outcomes—A scoping review”, *Behavioral Sciences*, MDPI, 1 September, doi: 10.3390/bs11090119.

Hyndman, R.J. and Athanasopoulos, G. (2021), *Forecasting: Principles and Practice*, 3rd ed., OTexts, Melbourne, Australia.

IBM. (2025a), “ What is a decision tree?”, available at: https://www.ibm.com/think/topics/decision-trees (accessed 26 June 2025).

IBM. (2025b), “ What is random forest? ”, available at: https://www.ibm.com/think/topics/random-forest (accessed 26 June 2025).

IBM. (2025c), “ What is XGBoost?”, *IBM*, available at: https://www.ibm.com/think/topics/xgboost (accessed 26 June 2025).

Kwiatkowski, D., Phillips, P.C.B., Schmidt, P. and Shin, Y. (1992), “Testing the null hypothesis of stationarity against the alternative of a unit root”, *Journal of Econometrics*, Vol. 54 No. 1–3, pp. 159–178, doi: 10.1016/0304-4076(92)90104-Y.

Luka, A. (2020), “Rolling and Expanding Windows For Dummies”, *Robot Wealth*, 25 May, available at: https://robotwealth.com/rolling-and-expanding-windows-for-dummies/ (accessed 28 September 2025).

Mathelinea, D., Chandrashekar, R. and Mawengkang, H. (2023), “Stationarity test for medicine time series data”, p. 030049, doi: 10.1063/5.0128444.

McCracken, M. and Ng, S. (2020), *FRED-QD: A Quarterly Database for Macroeconomic Research*, Cambridge, MA, doi: 10.3386/w26872.

McCracken, M.W. and Ng, S. (2021), “FRED-QD: A Quarterly Database for Macroeconomic Research”, *Review*, Vol. 103 No. 1, doi: 10.20955/r.103.1-44.

National Bureau of Economic Research. (2024), “Business Cycle Dating Procedure: Frequently Asked Questions”, 23 September, available at: https://www.nber.org/research/business-cycle-dating/business-cycle-dating-procedure-frequently-asked-questions (accessed 29 December 2024).

NBER. (2023), “US Business Cycle Expansions and Contractions”, 14 March, available at: https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions (accessed 29 December 2024).

Nyman, R. and Ormerod, P. (2017), “Predicting Economic Recessions Using Machine Learning Algorithms”.

Nyman, R. and Ormerod, P. (2020), “Understanding the Great Recession Using Machine Learning Algorithms”.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., *et al.* (2021), “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews”, *BMJ*, p. n71, doi: 10.1136/bmj.n71.

Petropoulos, A., Siakoulis, V., Panousis, K.P., Papadoulas, L. and Chatzis, S. (2023), “Macroeconomic forecasting and sovereign risk assessment using deep learning techniques”.

Pontes, E.L., Benjannet, M. and Yung, R. (2024), “Forecasting Four Business Cycle Phases Using Machine Learning: A Case Study of US and EuroZone”.

Qilu, Y. (2022), *MACHINE LEARNING APPLICATIONS IN ECONOMICS*, Cornell University, May.

Rodrigo, J.A. and Ortiz, J.E. (2024a), “Introduction to forecasting”, *Skforecast Docs*.

Rodrigo, J.A. and Ortiz, J.E. (2024b), “Exogenous variables (features)”, available at: https://skforecast.org/0.12.1/user_guides/exogenous-variables.html (accessed 20 September 2025).

Shailesh. (2024), “ Various Techniques to Detect and Isolate Time Series Components Using Python”, *Analytics Vidhya*, 18 May, available at: https://www.analyticsvidhya.com/blog/2023/02/various-techniques-to-detect-and-isolate-time-series-components-using-python/ (accessed 15 June 2025).

Sties, M. (2017), *Forecasting Recessions in a Big Data Environment*.

Tin Kam Ho. (1998), “The random subspace method for constructing decision forests”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20 No. 8, pp. 832–844, doi: 10.1109/34.709601.

Wang, T., Beard, R., Hawkins, J. and Chandra, R. (2024), “Recursive Deep Learning Framework for Forecasting the Decadal World Economic Outlook”, *IEEE Access, Access, IEEE*, IEEE, Vol. 12, pp. 152921–152944, doi: 10.1109/ACCESS.2024.3472859.

Weinstock, L.R. (2023), *Common Causes of Economic Recession* .

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X. and Xu, H. (2022), “Time Series Data Augmentation for Deep Learning: A Survey”, doi: 10.24963/ijcai.2021/631.

"What Is Principal Component Analysis (PCA)? | IBM". (n.d.). , available at: https://www.ibm.com/think/topics/principal-component-analysis (accessed 1 September 2025).

xgboost developers. (2022), "Introduction to Boosted Trees", available at: https://xgboost.readthedocs.io/en/stable/tutorials/model.html (accessed 26 June 2025).

Zhou, Z.-H. (2012), *Ensemble Methods*, Chapman and Hall/CRC, doi: 10.1201/b12207.

Zyatkov, N. and Krivorotko, O. (2021a), "Forecasting Recessions in the US Economy Using Machine Learning Methods", *2021 17th International Asian School-Seminar "Optimization Problems of Complex Systems (OPCS), School-Seminar "Optimization Problems of Complex Systems (OPCS), 2021 17th International Asian*, IEEE, 13 September, doi: 10.1109/OPCS53376.2021.9588678.

Zyatkov, N. and Krivorotko, O. (2021b), "Forecasting Recessions in the US Economy Using Machine Learning Methods", *2021 17th International Asian School-Seminar "Optimization Problems of Complex Systems (OPCS)*, IEEE, pp. 139–146, doi: 10.1109/OPCS53376.2021.9588678.

Appendix A

In this appendix, the artifacts produced in the scope of the “Data Preparation and Modelling” chapter are presented.

Table 11 presents the list of variables from the dataset used as features in Feature Selection Strategy 1 (“All Variables”)

Table 12 presents the subset of features selected through Pearson Correlation Filtering, which were subsequently used for training, validation, and testing in the context of the Modelling phase.

Table 11 - List of variables from dataset used as features in Feature Selection Strategy 1 (“All Variables”)

PCECC96	PCESVx	A014RE1Q156NBEA	GCEC1
PCDGx	PCNDx	GPDIC1	A823RL1Q225SBEA
FPIx	Y033RC1Q027SBEAx	PNFIx	PRFIx
FGRECPTx	SLCEx	EXPGSC1	IMPGSC1
DPIC96	OUTNFB	OUTBS	INDPRO
IPFINAL	IPCONGD	IPMAT	IPDMAT
IPNMAT	IPDCONGD	IPB51110SQ	IPNCONGD
IPBUSEQ	IPB51220SQ	CUMFNS	PAYEMS
USPRIV	MANEMP	SRVPRD	USGOOD
DMANEMP	NDMANEMP	USCONS	USEHS
USFIRE	USINFO	USPBS	USLAH
USSERV	USMINE	USTPU	USGOVT
USTRADE	USWTRADE	CES9091000001	CES9092000001
CES9093000001	CE16OV	CIVPART	UNRATE
UNRATESTx	UNRATELTx	LNS14000012	LNS14000025
LNS14000026	UEMPLT5	UEMP5TO14	UEMP15T26
UEMP27OV	LNS12032194	HOABS	HOANBS
AWHMAN	AWOTMAN	HWIx	HOUST
HOUST5F	HOUSTMW	HOUSTNE	HOUSTS
HOUSTW	AMDMNOx	AMDMUOx	PCECTPI
PCEPILFE	GDPCTPI	GPDICTPI	IPDBS
DGDSRG3Q086SBEA	DDURRG3Q086SBEA	DSERRG3Q086SBEA	DNDGRG3Q086SBEA
DHCERG3Q086SBEA	DMOTRG3Q086SBEA	DFDHRG3Q086SBEA	DREQRG3Q086SBEA
DODGRG3Q086SBEA	DFXARG3Q086SBEA	DCLORG3Q086SBEA	DGOERG3Q086SBEA
DONGRG3Q086SBEA	DHUTRG3Q086SBEA	DHLCRG3Q086SBEA	DTRSRG3Q086SBEA
DRCARG3Q086SBEA	DFSARG3Q086SBEA	DIFSRG3Q086SBEA	DOTSRG3Q086SBEA
CPIAUCSL	CPILFESL	WPSFD49207	PPIACO
WPSFD49502	WPSFD4111	PPIIDC	WPSID61
WPU0561	CES2000000008x	CES3000000008x	COMPRNFB
RCPHBS	OPHNFB	OPHPBS	ULCBS
ULCNFB	UNLPNBS	FEDFUNDS	TB3MS
TB6MS	GS1	GS10	AAA
BAA	BAA10YM	TB6M3Mx	GS1TB3Mx
GS10TB3Mx	BOGMBASEREALx	M1REAL	M2REAL
BUSLOANSx	CONSUMERx	NONREVSLx	REALLNx
TOTALSLx	EXSZUSx	EXJPUSx	EXUSUKx
EXCAUSx	B020RE1Q156NBEA	B021RE1Q156NBEA	IPMANSICS
IPB51222S	IPFUELS	UEMPMEAN	CES0600000007
TOTRESNS	NONBORRES	GS5	TB3SMFFM

T5YFFM	AAAFFM	WPSID62	PPICMM
CPIAPPSL	CPITRNSL	CPIMEDSL	CUSR0000SAC
CUSR0000SAD	CUSR0000SAS	CPIULFSL	CUSR0000SA0L2
CUSR0000SA0L5	CES0600000008	DTCOLNVHFNM	DTCTHFNM
INVEST	HWIURATIOx	BUSINVx	ISRATIOx
CONSPIx	NIKKEI225	CNCFx	S&P 500
S&P div yield	S&P PE ratio		

Table 12 - Features chosen through Pearson Correlation Filtering (Strategy 2) for use in training, validation, and testing phases.

PCECC96	OUTBS	OUTNFB	PCESVx
DRCARG3Q086SBEA	USEHS	DPIC96	PCNDx
CUSR0000SAS	DHCERG3Q086SBEA	DSERRG3Q086SBEA	OPHPBS
OPHNFB	DOTSRG3Q086SBEA	USPBS	CES0600000008
DHUTRG3Q086SBEA	CPIMEDSL	DFSARG3Q086SBEA	CPIAUCSL
CPIULFSL	DHLCRG3Q086SBEA	CPILFESL	CUSR0000SA0L5
FGRECPTx	TOTALSLx	BUSINVx	CUSR0000SA0L2
FPIx	GDPCTPI	DFXARG3Q086SBEA	UNLPNBS
PCECTPI	GPDIC1	DIFSRG3Q086SBEA	IMPGSC1
PCEPILFE	IPMAT	DTRSRG3Q086SBEA	COMPRNFB
USLAH	EXPGSC1	WPSFD49502	PNFIx
RCPHBS	WPSFD49207	WPSFD4111	CPITRNSL
DONGRG3Q086SBEA	GCEC1	SRVPRD	Y033RC1Q027SBEAx
IPDBS	DNDGRG3Q086SBEA	PPIIDC	PPIACO
CNCFx	USPRIV	CONSUMERx	REALLNx
WPSID61	CUSR0000SAC	PAYEMS	IPDMAT
INDPRO	CE16OV	SLCEx	IPBUSEQ
ULCNFB	ULCBS	NONREVSLx	DTCOLNVHFNM
BUSLOANSx	IPB51110SQ	DTCTHFNM	HOABS
PCDGx	IPB51220SQ	HOANBS	IPMANSICS
USFIRE	IPB51222S	IPDCONGD	USSERV
IPFINAL	USCONS	GPDICTPI	M2REAL
USTPU	CES9093000001	IPFUELS	INVEST
DMOTRG3Q086SBEA	WPSID62	DGOERG3Q086SBEA	USGOVT
USTRADE	DGDSRG3Q086SBEA	B021RE1Q156NBEA	CES9092000001
IPCONGD	CES3000000008x		