Methodology for GDP Forecasting Using Machine Learning

1. Introduction

This report outlines the methodology used to forecast GDP for the next five quarters using a Linear Auto-Regression model. The approach includes data preprocessing, model training, prediction, and visualization to analyze economic trends.

2. Data Preprocessing

2.1 Data Selection and Splitting

The `Quarterly Data.csv` dataset consists of multiple economic indicators, including:

- **Gross Domestic Product (GDPC1)**, **Real Government Consumption Expenditure & Gross Investment (GCEC1), Real Exports of Goods & Services (EXPGSC1)**
- **Volatility Index (VIX), S&P 500 Index, University of Michigan Consumer Sentiment Index (UMC)**

The dataset was divided into training (**df_train**) and testing (**df_test**) sets:

- **df_train**: First 174 observations (70%)
- **df_test**: Remaining 30% of the dataset

Relevant indicators were selected for their direct impact on GDP through various economic approaches:

Variable Code	Rationale
GDPC1	It represents inflation-adjusted GDP, making it a reliable indicator of economic growth.
PCECC96	It has a direct causal relationship with Real GDP through the expenditure approach.
GDPIC1	It has a direct causal relationship with Real GDP through the expenditure approach.
GCEC1	It has a direct causal relationship with Real GDP through the expenditure approach.
EXPGSC1	It has a direct causal relationship with Real GDP through the expenditure approach.
IMPGSC1	It has a direct causal relationship with Real GDP through the expenditure approach.
FEDFUNDS	Nominal Federal rates will affect Real GDP directly through cost of borrowing and investment.
CPIAUSCL	It measures inflation which directly affects Real GDP.

2.1.1 Data Interpretation

2.1.1.1 Expenditure Approach for Real GDP Calculation

The **Expenditure Approach** is a well known method to calculate **Real GDP**, based on the total spending on goods and services in an economy. This approach focuses on demand side economic activity and it is easy to measure. The formula is:

$$\textit{Real GDP} = \textit{Consumption} + \textit{Investment} + \textit{Government Expenditure} + (\textit{Exports} - \textit{Imports})$$

2.1.1.2 Inflation Rate Approach for Real GDP Calculation

Inflation directly affects Real GDP because it influences the purchasing power of money and the true value of economic output. Hence, we are interested in using inflation rate to calculate Real GDP. However, in FRED-QD, inflation rate was not provided directly for our calculation. As a result, we decided to make use of CPIAUSCL to calculate inflation rate. This was presented by calculating the percentage change in CPIAUSCL across the quarters.

$$\textit{Inflation Rate} = \textit{(CPI in Current Year} - \textit{CPI in Previous Year) / CPI in Previous Year * 100}$$

After obtaining the inflation rate, we are able to calculate real interest rates using the Fisher Equation. We believe that real interest rates help predict economic downturns by showing the true cost of borrowing and the real return on savings after adjusting for inflation. The formula is:

$$\textit{Real Interest Rate} = \textit{(1+Inflation Rate)/(1+Nominal Interest Rate)} - 1$$

2.2 Data Exploration

2.2.1 STATA: Linear Regression

We conducted statistical analysis using STATA with the `quarterly_perc.csv` dataset. In this analysis, we fitted a linear regression model with the components of the expenditure approach and the `gdp_change` as the dependent variable. It's important to note that all values in the `quarterly_perc.csv` file are expressed as percentage changes. These values are calculated by determining the difference between the current period's value and the previous period's value, then converting the result into a percentage.

The linear regression model yielded a high R-squared value of 0.9908, indicating that the model explains a significant portion of the variance in the data. The Root Mean Square Error (RMSE) value is 0.10204, which further supports the model's accuracy.

The fitted linear equation is as follows:

$$\begin{aligned}\textbf{GDP} = {} & 0.6324521 \cdot \text{consump\_change} + 0.1676679 \cdot \text{invest\_change} + 0.2346663 \cdot \text{govt\_change} + \\ & 0.0805634 \cdot \text{exports\_change} - 0.0866884 \cdot \text{imports\_change} - 0.0289958\end{aligned}$$

Unfortunately, we were unable to forecast GDP for Q1, Q2, and Q4 of 2025 due to the lack of data for the future periods. However, this regression model will serve as the foundational concept for developing our Linear Auto-Regression Model in the future.

2.2.2 STATA: ARIMA

Next, we attempted to implement an ARIMA model, a statistical analysis technique that uses time series data to forecast future trends. To set up the model, we needed to configure its parameters. Using the `dfuller` test, we found that the probability was 0, indicating that the data is stationary, so we set d=0.

Next, we examined the autocorrelation function (ACF) and partial autocorrelation function (PACF) using the `corrgram`. The PACF did not exhibit a clear sharp cutoff at lower lags, suggesting a lower value for p. Meanwhile, the ACF showed small values across most lags, with no strong peaks, pointing to a small value for q.

With this information in hand, we manually adjusted the values for p and q, comparing the resulting p-values for the chi-square test (Prob > chi2). Although we initially considered using the AIC/BIC method, which suggested an AR(2) model, the changes in AIC/BIC across different lags were minimal, so we decided to focus on comparing Prob > chi2.

Ultimately, we settled on an ARIMA model with the specification **arima gdp_change, ar(3) ma(1)**. The resulting Prob > chi2 value was 0.7641, indicating a reasonably good fit.

However, due to our limited experience with STATA, we encountered difficulties in forecasting the future values despite multiple attempts and adjustments.

2.3 Feature Engineering

To incorporate temporal dependencies, sequences of length **10** were created:

- **create_sequences(data, seq_length)**: generates sequences for independent variables
- **create_sequences_y(data, seq_length)**: generates target variable sequences (GDP)

Each economic indicator was converted into a sequence format before being used as input.

3. Model Training and Prediction (Linear Auto-Regression and LSTM)

3.1 Linear Auto-Regression

Linear Auto-Regression is a fundamental machine learning algorithm which uses predictive modeling, to establish a linear relationship between one or more independent variables and a dependent variable.

Linear Auto-Regression follows the equation of a straight line:
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$, where:

- Y = Dependent Variable
- $X_1, X_2, \ldots, X_n$ = Independent Variable
- β_0 = Intercept (Value of Y when X=0)
- $\beta_1, \beta_2, \ldots, \beta_n$ = Coefficients that determine the impact of each independent variable
- ϵ = Error term, representing the difference between actual and predicted values

The objective is to estimate the coefficient (B) that fits the data using Ordinary Least Squares (OLS), which minimises the sum of squared errors.

3.1.1 Data Preparation

We first trained the model to learn the relationship between input features and the target variable by determining the best-fit coefficients. This is achieved by minimizing the difference between actual and predicted values. After the model is trained, it is then used to predict future values of Y based on the input of X, which are the various factors mentioned above. Following which, we evaluate the performance of the predictions done by the trained model, by calculating the Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

To decide on the time period, t, to run the data back by, we ran the trained Linear Auto-Regression Model through an iteration with t having a range of 1 to 20, From there, we found the smallest RMSE to conclude the value of t which we will be using, which is 1. The table below shows the results of the RMSE collected.

Time sequence input	RMSE
1	2.7539
2	2.8581
3	4.7727

3.1.2 Linear Auto-Regression Results

Following which, we ran the Linear Auto-Regression model with t=1, and got the following predictions for the next 5 quarters after the last data set given. The table below shows the predicted results.

Quarter	Predicted GDP percentage change (%)
Q4 2024	0.89
Q1 2025	1.06
Q2 2025	0.29
Q3 2025	-0.62
Q4 2025	-1.38

3.2 Long Short-Term Memory (LSTM)

Long short-term memory (LSTM) is a machine learning model that makes use of an algorithm called neural network (NN) aimed at mitigating the vanishing gradient problem commonly encountered by traditional recurrent NNs. LSTMs are great for time series forecasting because they capture long-term dependencies, recognize complex patterns, and can handle multiple input features. They excel at learning trends and seasonality, making them useful for economic predictions, but they require large datasets and are less interpretable than traditional models.

3.2.1 Data Preparation

Given the train and test data, one has to first figure out the time period t as to how far back to train the data on. Similar to the Linear Auto-Regression model above, we iterated through values of time from 1 to 20 and ran the LSTM model to identify the value of time that gives the smallest root mean squared error (RMSE). The results are in the table below, and the lowest RMSE was achieved at a time value of 6.

Time sequence input	RMSE
5	161.00
6	45.32
7	264.04

3.2.2 LTSM Results

Using the derived time value of 16, we then ran the LTSM algorithm to predict the following 5 quarters (in order to make predictions for Q4 2025). The following results were derived.

Quarter	Predicted GDP percentage change (%)
Q4 2024	-1.50
Q1 2025	-2.22
Q2 2025	-2.41
Q3 2025	-2.55
Q4 2025	-2.49

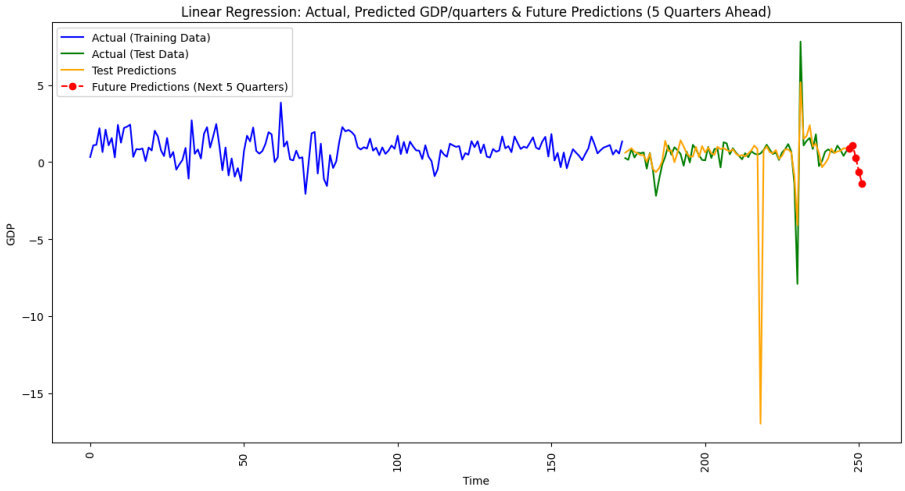
3.3 Model Selection

Overall, we looked at the root mean squared error (RSME) as the main indicator of model reliability. Further, the LTSM had an RMSE close to 20 times larger than that of the linear auto-regression model, thus showing the inconsistent nature of the former model. As such, the linear auto-regression model was chosen due to its high performance in test data, as well as its simplicity and interpretability.

4. Overall Results

4.1. Visualization and Analysis

A time-series plot was generated to visualize:



This graph helped to visualise the model's performance and trends in GDP forecasting, based on the various inputs.

5. Summary

5.1 Key Findings

As seen in section 3.1.2, the economy is predicted to continue to perform well in Q1 and Q2 2025. However, a downturn is predicted to take place in Q4 2025, with a change in GDP of -1.38%. As such, major market players should be more conservative when looking to invest in the last quarter, and err on the side of caution.

We also took an additional step to identify the key indicators in our model that significantly drive changes in GDP. With the weights derived from our model, we established a regression equation for us to visualise the impact of the indicators.

gdp_change = 3.20e-01consump_change - 3.47e-02invest_change - 1.21e-01govt_change - 2.91e-02exports_change + 8.27e-02imports_change - 2.65e-03vix_change + 3.05e-02sp500_change - 5.24umc_senti_change + 9.95e-05inflation_change - 1.56e-03nominal_int_change + 7.82e-04real_int_change

As seen from the above equation, the indicators with the greatest impact, indicated by the larger coefficients, are *consumption_change, invest_change, govt_change, exports_change, imports_change*. This further highlights how important it is to monitor these components when it comes to predicting changes in GDP.

5.2 Limitations

This study has several limitations that could impact the reliability of the results. First, the cleaned `quarterly_perc.csv` dataset may not be sufficient for more complex models like LSTM, which typically require large amounts of data to perform effectively. The reduction in the number of variables in the dataset may have led to the exclusion of factors that could have a significant impact on `gdp_change`.
Furthermore, LSTM models are often considered “black box” models, meaning that the coefficients and predictions may vary with each run, which can lead to inconsistencies and errors in forecasting as the model's results are not fixed.
Additionally, the assumptions made in the calculation of inflation and real interest rates could introduce measurement errors, which might further compromise the accuracy of the findings.

In conclusion, we decided to go with the forecast results from the linear auto-regression model. According to these predictions, we’re expecting an economic downturn in Q4 2025, with a contraction of -1.38%. However, in Q1 and Q2 2025, the economy is expected to see positive growth, with GDP increasing by 1.06% and 0.29%, respectively.