Forecasting US GDP Growth Rate

1 Introduction

Accurate forecasting of GDP growth is crucial for policymakers, investors and economic analysts, as it provides valuable insights into the future of the state of the economy. Predicting US GDP is especially relevant for Singapore due to the close economical and financial ties between the 2 countries, as well as the significant influence of the US economy on the global economic landscape. Given the complexity and uncertainty inherent in economic systems, reliable models for predicting GDP growth are essential in making policy decisions, investment strategies, and economic planning.

Traditional models such as ARIMA and vector autoregression (VAR) have been widely used for economic forecasting due to their simplicity and interpretability (Hamilton, 1994; Stock & Watson, 1999). However, with advancements in data science and machine learning, there is interest in exploring newer techniques to improve forecast accuracy and better understand economic dynamics. New methods such as Random Forest and LASSO are introduced in frequent years, offering flexible, data-driven approaches for capturing complex patterns (Athey & Imbens, 2019; Varian, 2014). By building on both traditional and modern methods,

our report aims to contribute to the literature on forecasting the Industrial Production Index (INDPRO) with a comprehensive analysis of model performance across horizons and integration of hybrid methods that capture both linear and nonlinear dynamics.

2 Data

2.1 Dataset

In alignment with our problem statement, we chose the FRED-MD dataset, which includes a broad spectrum of economic variables from sectors like the labour market, money and credit, stock market, etc. The dataset frequency is monthly, spanning from 01-1959 to 08-2024.

The span of the dataset we chose to analyse is 03-1959 to 12-2019. We did not choose the whole span of the dataset due to the obvious 15% change outlier (refer to *Fig2*) we noticed in 2020-2021, which might be due to the occurrence of COVID-19.

We have opted to use variables from the FRED-MD database. In our data, we have a sample of 787 observations and 126 variables with 23 missing observations. We considered 4 lags of each variable. Overall, our analysis consists of 504 potential predictors.

The dataset provided was in the form of an untransformed time series. This initial visualisation revealed distinct economic cycles, including identifiable periods of recession and growth.

2.2 Target Variable

In this project, we decided to use the Industrial Production Index (INDPRO) as a proxy for the US gross domestic product (GDP). INDPRO measures the real output of manufacturing, mining and utilities in the US sectors that significantly influence GDP fluctuations. While GDP encompasses the entire economy, including services, INDPRO provides a more immediate snapshot of industrial performance, which is often a leading indicator of broader economic trends.

2.3 Data Transformation

For Data Cleaning, we omitted 23 variables that have missing values entirely from the dataset, resulting in a final dataset consisting of 103 indicators and 412 potential indicators, with 4 lags of each variable. Since time series analysis often requires stationarity for accurate modelling, we followed the transformation guidelines outlined in the St. Louis Fed working paper (2015-012). These transformations were applied to achieve stationarity across variables, preparing the dataset for reliable predictive modelling.

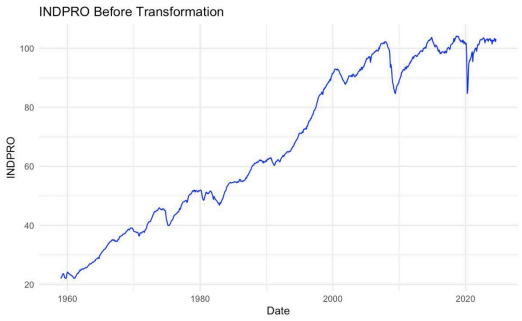

Fig 1: Time Series Of INDPRO before Transformation

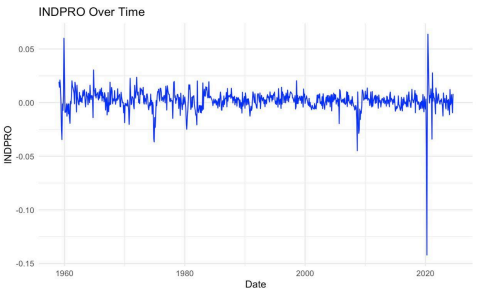

Fig 2: Time Series Of INDPRO after Transformation

Through data analysis, we identified the outlier to occur in 04-2020. Hence decided to only use the time period up till 12-2019 with 730 observations for further forecasting analysis.

```
##      sasdate     INDPRO
## 1 2020-04-01 -0.1420453
```

Fig 3: Results of outlier analysis

2.4 Exploratory Data Analysis

2.4.1 Correlation Analysis

A correlation matrix was generated to assess multicollinearity among predictors. This analysis revealed that multiple pairs of predictors exhibit high correlations (using the benchmark of absolute correlation > 0.7). In total, 334 predictor pairs were identified with significant correlations, indicating a strong presence of multicollinearity in the dataset.

High multicollinearity suggests that many predictors contain insignificant information which may complicate the model's stability and results. Given these findings, we hypothesise that elastic net could be advantageous over LASSO for regression-based modelling as it is more effective in handling groups of correlated predictors.

2.4.2 High Variance Analysis

To further understand the characteristics of the predictors, variance analysis was conducted, focusing on the top 25% highest variance predictors. High variance predictors often contain valuable information but may also contain noise, which affects the model's performance. Our analysis shows high variance amongst predictors, suggesting that methods like bagging or

random forest could be beneficial as they reduce variance and are capable of capturing complex patterns in high-variance datasets.

2.4.2 Pairwise Scatter Plots of High Variance Predictors

To explore potential relationships and interactions among these high-variance predictors, pairwise scatterplots were generated. The plots suggested that most predictor pairs do not exhibit clear linear relationships and instead display scattered or non-linear relationships. In addition, a few outliers are visible. The presence of linear relationships further suggests the suitability of bagging and random forest for this dataset.

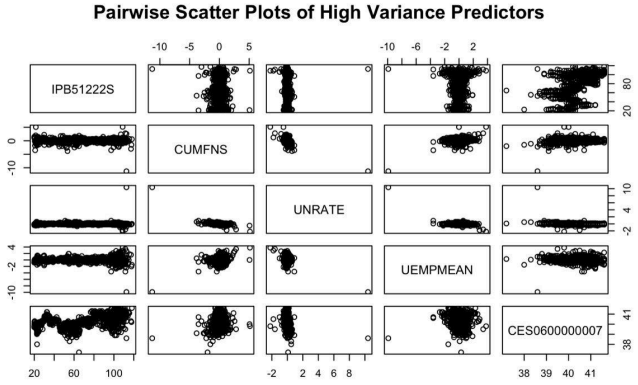

Fig 4: Pairwise scatterplot of leading variables

2.5 Time Horizon

We chose 1-, 3-, 6- and 12- step forecast horizons to balance short-term and long-term perspectives on economic forecasting. This approach aligns with recent research, such as Zhang et al. (2023), who emphasised the importance of benchmarking forecasting models across diverse prediction horizons. Their study demonstrates that evaluating models over different horizons enhances the robustness and adaptability of forecasts, providing insights for both immediate and extended periods.

3 Methodology

To evaluate the forecasting performance of the various machine learning methods, an Autoregressive (AR) model with p lags was used as a benchmark since it is a simple model that can capture the persistence in the transformed, stationary data. The number of p lags was determined by selecting the model with the lowest Bayesian Information Criterion (BIC) for a more parsimonious fit since it is more consistent and tends to select the true model as the sample size grows. The machine learning models implemented with the HDeconometrics package included penalised regression methods, such as LASSO, with its lambda chosen by BIC, Akaike Information Criteria (AIC), and corrected AIC (AICc), post-LASSO using rlasso,

and elastic net with an alpha of 0.5 and its lambda chosen by BIC. These sparsity-inducing shrinkage models shrink coefficients of less relevant variables to 0, potentially improving model interpretability and overfitting since we have a large number of predictors. We also applied ensemble learning models such as bagging, and a random forest model on default settings, to capture non-linear patterns in the data and aid in model selection. Additionally, we also consider a forecast combination of the best 5 models by averaging over the models using a simple average, to smooth over the individual model weaknesses and integrate their strengths.

To train our models, the data was split into test and training datasets, where the test set contained one-third of the total number of observations, amounting to 243 observations from September 1999 to December 2019. Direct 1-step, 3-step, 6-step and 12-step ahead forecasts were generated using a rolling estimation window of fixed length of 487 observations to ensure that forecasts were made using the most recent and relevant data available that remains covariance stationary. The root mean squared error (RMSE) values of the forecasts under the squared loss function were then computed for each of the models to compare their performance. These RMSE values will be multiplied by 100 for ease of interpretation. In addition, Diebold-Mariano (DM) tests were conducted on the models with the lowest RMSE values to determine whether there is a significant difference in their predictive abilities.

4 Empirical Results

4.1 Benchmark: AR(4)

The AR(4) model was chosen as our benchmark, with the lowest BIC of -9.9594. The table below shows the RMSE values of each forecast horizon, with the values increasing with the forecast horizon lengths. This could be due to the 4 lags of INDPRO losing relevance in predicting longer-term dynamics as other predictors, such as the labour market indicators that represent supply-side factors of production, become more significant at longer forecast horizons.

Forecast Horizon	RMSE (multiplied by 100)

1-step	0.6199083
3-step	0.6249955
6-step	0.6655774
12-step	0.6775993

Fig 5: AR(4) h-step ahead RMSEs multiplied by 100

The coefficients of the AR model exhibit minor fluctuations, but still remain relatively stable which suggests that the model is generally consistent in capturing the dynamics of the data over time. The AR(4) model follows the general patterns of the data in the short term quite closely, but its performance worsens significantly in the long term where the lags become less effective in explaining the variation in the data, which is evident in the increase in RMSE with longer forecast horizons.

4.2 Penalised Regressions: LASSO Variants

The following RMSE values were obtained for the LASSO models, with the values multiplied by 100 for ease of interpretation and the model with the best performance highlighted in bold. It is important to note that the models with lambda chosen on AIC and AICc with small sample correction perform the worst, since AIC has a tendency to overfit because it has a smaller penalty on the number of parameters. Therefore, we chose to select the lambdas for elastic net and post-LASSO using BIC. Firstly, LASSO (BIC) has a lower RMSE than the AR(4) benchmark, but only differing slightly by 0.0005. For the 3- and 6- step ahead forecasts, the benchmark outperforms the LASSO variants since LASSO tends to select only one variable in a group of highly correlated variables. However, in the 12-step ahead forecast, the post-LASSO model has the best performance with the lowest RMSE of 0.661, because it undoes the shrinkage bias by applying ordinary least squares (OLS) estimations to the predictors selected by LASSO (Belloni & Chernozhukov, 2013). This helps to generate more accurate and stable predictions, especially in the long term, since the predictors are highly correlated.

Forecast	AR(4)	LASSO	LASSO	LASSO	Elastic Net	Post-LASSO

Horizon		(BIC)	(AIC)	(AICc)	(BIC, α = 0.5)	(BIC, α = 1)
1-step	0.6199083	**0.6194118**	7.073549	0.6236486	0.6332132	0.6226274
3-step	**0.6249955**	0.6665512	3.252365	0.6553819	0.6791659	0.6479973
6-step	**0.6655774**	0.6808682	8.114531	0.8380212	0.6754248	0.7144949
12-step	0.6775993	0.6613267	7.499521	1.684541	0.6616369	**0.6553784**

Fig 6: Penalised Regressions vs AR(4) h-step ahead RMSEs multiplied by 100

Examining the performance of the LASSO models alone, the LASSO (BIC), elastic net and post-LASSO models have comparable performance, with a maximum difference of approximately 0.03. We therefore select these models for our forecast combinations and Diebold-Mariano tests.

As seen in the sparsity analyses for both LASSO (BIC) and elastic net in the figures below, there has been a significant amount of variable selection which generally appears to be parsimonious over time, with a range of 2 to 25 predictors out of the 412 predictors across the 4 forecast horizons. Although the elastic net model is expected to select more variables than LASSO because of the ridge component that shrinks correlated predictors together, it selects a similar number of variables to the LASSO model in this case. This might be due to the lack of correlated groups of variables that are useful in predicting INDPRO. It can also be observed that for both models, the number of variables selected by the models decreases as the forecast horizon lengthens because the model is unable to identify variables with significant relevance for predicting values in the distant future.

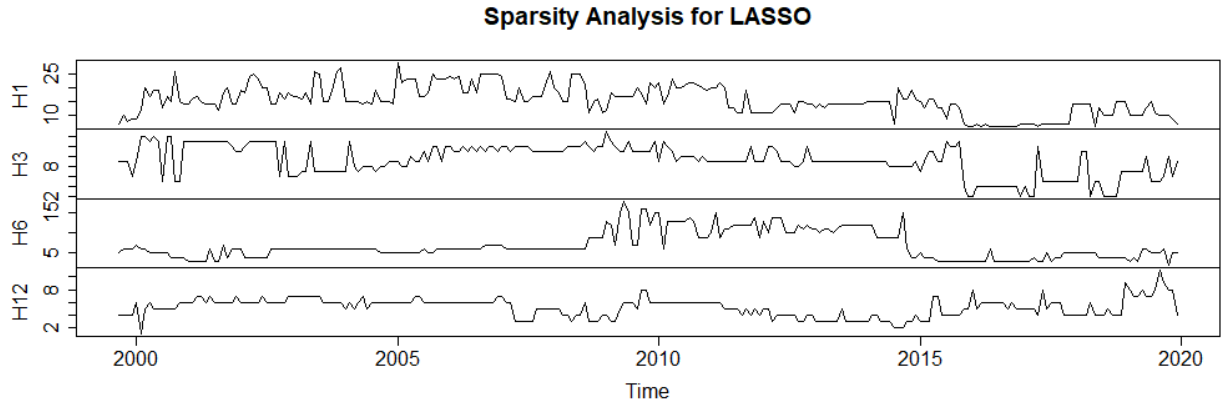

Fig 7: Sparsity Analysis for LASSO (BIC)

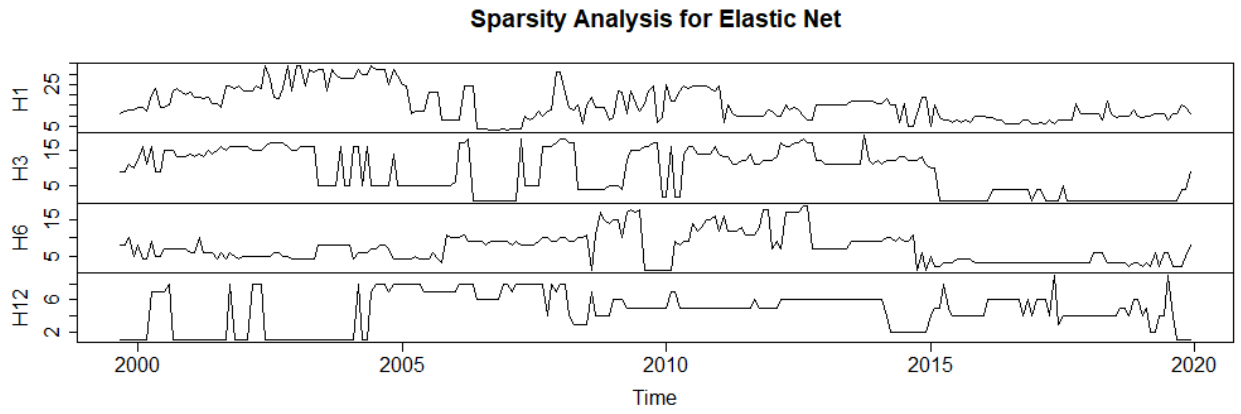

Fig 8: Sparsity Analysis for Elastic Net (BIC)

4.3 Bagging and Random Forest

4.3.1 Model Specifications

Both Bagging and Random Forest were done on default settings of letting the tree grow to 5 observations per node and B = 500. For bagging, the number of predictors = P, which was 103, while for random forest, the number of predictors = P/3, which was about 34.

Rolling window estimation was used, and the testing dataset was fixed to be one-third of the total dataset, amounting to 243 observations.

4.3.2 Results and Visualisation

4.3.2.1 Bagging

Below are the RMSE values for each of the models for each forecast horizon. The models have very similar RMSE values even as the forecast horizon increases. Low RMSE values indicate that bagging provides a good model for prediction of US GDP.

Forecast Horizon	RMSE (multiplied by 100)
1-Step	0.1829557
3-Step	0.1887247
6-Step	0.1911559
12-Step	0.1903657

Fig 9: Bagging h-step ahead RMSEs multiplied by 100

The plots of the actual (blue line) against the predicted values (red dotted line) support the good fit of the Bagging models for each forecast horizon.

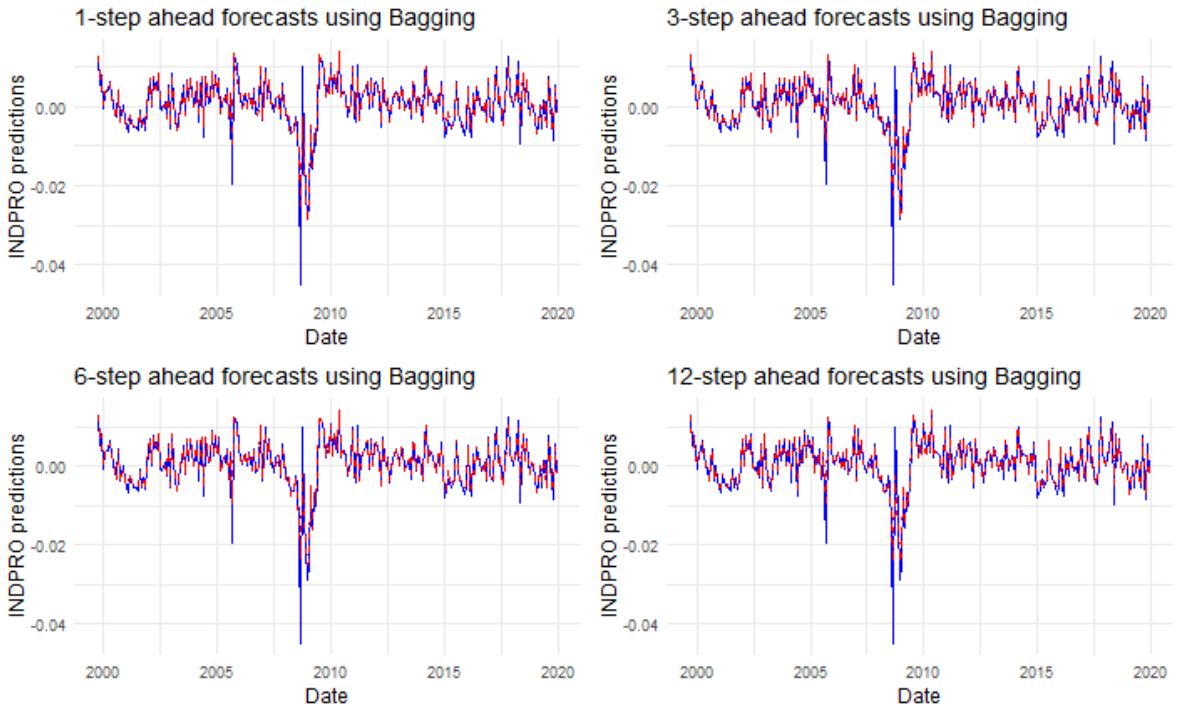

Fig 10: Visualisations of predictions for each Bagging model for each forecast horizon

4.3.2.2 Random Forest

Random Forest seems to work similarly to Bagging, with Bagging outperforming Random Forest by a small margin. Random Forest still has a good performance with low RMSE values as seen below.

Forecast Horizon	RMSE (multiplied by 100)
1-Step	0.2125833
3-Step	0.2162928
6-Step	0.2155475
12-Step	0.2146593

Fig 11: Random Forest h-step ahead RMSEs multiplied by 100

The plots of the actual (blue line) against the predicted values (red dotted line) support the good fit of the Random Forest models for each forecast horizon.

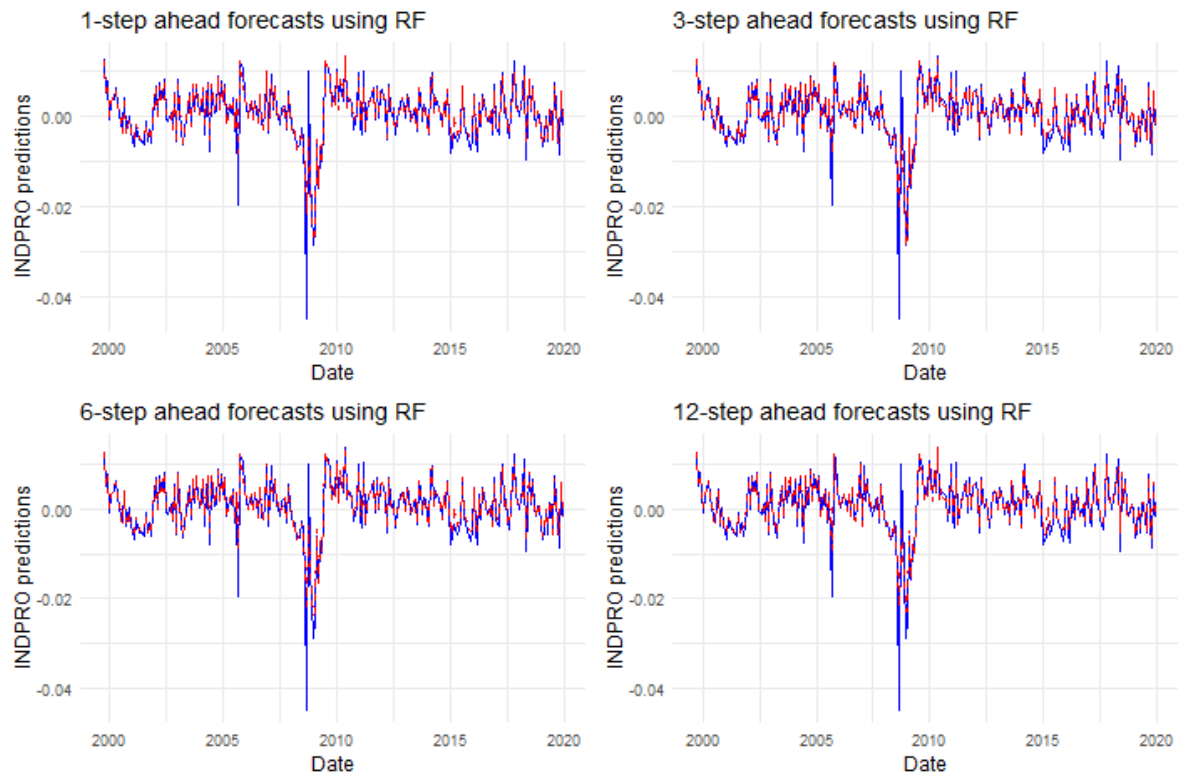

Fig 12: Visualisations of predictions for each Random Forest model for each forecast horizon

The Variable Importance Plots were also plotted, where the variables were ranked based on their improvement to the model at each split. As seen from the plots below, no matter the forecast horizon, IPMAT (IP: Materials), is the most important variable, adding approximately 25% value to each model. Among all the models for each forecast horizon, all of the variables repeatedly appear in all 4 plots. Aside from IPMAT, the rest are CUMFNS, IPDMAT, IPCONGD, IPFPNSS, IPMANSICS, IPFINAL, IPNMAT and IPBUSEQ. These variables are Capacity Utilization: Manufacturing, IP: Durable Materials, IP: Consumer Goods, IP: Final Products and Nonindustrial Supplies, IP: Manufacturing (SIC), IP: Final Products (Market Group), IP: Nondurable Materials and IP: Business Equipment respectively. This shows that variables indicating industrial production (IP) could be more useful in predicting US GDP, and that the same variables are important to forecasting US GDP in each forecast horizon.

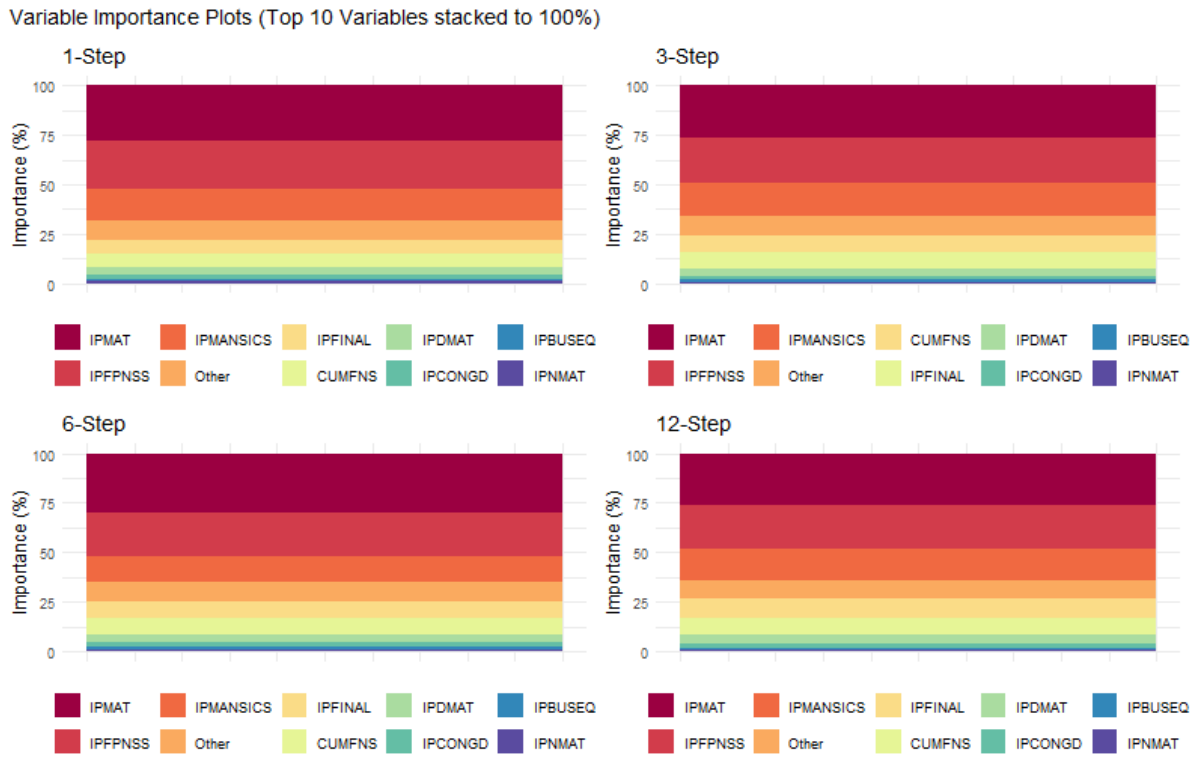

Fig 13: Random Forest Variable Importance Plots for each forecast horizon

4.4 Ensemble: Simple Average

To construct our ensemble model, we took a simple average combination of predictions from the AR(4), LASSO (BIC), elastic net, post-LASSO and random forest models. The simple average model with Random Forest performs much better than the AR(4) model. This could be due to the fact that combining LASSO and random forest can not only extract sparse linear signals, but also detect nonlinearities and interactions in the data.

Due to the significantly lower RMSE values of the Random Forest model, the resulting RMSE of the Simple Average model is also much lower than AR(4). Therefore, we have also included a model without the Random Forest. Other than the 3-step and 6-step ahead forecasts, the simple average model without random forest has slightly lower RMSE values than the AR(4) benchmark. This suggests that a simple average of forecast combinations can outperform models in the shorter and longer horizons, while losing out in the middle length horizons, as it leverages on each model's strengths and smooths over their weaknesses.

Forecast Horizon	AR(4)	Simple Average w RF	Simple Average w/o RF
1-Step	0.6199083	**0.520248**	**0.6150084**
3-Step	0.6249955	**0.5443755**	0.6467901
6-Step	0.6655774	**0.5648872**	0.6737073
12-Step	0.6775993	**0.5569487**	**0.659678**

Fig 14: Simple Average vs AR(4) h-step ahead RMSEs multiplied by 100

4.5 Diebold-Mariano Test

To test whether the benchmark, chosen LASSO variants, bagging and random forest have equal predictive ability, we use the DM equal predictive ability test for pairwise combinations of the models. Since the test set with a size of 243 observations is large enough, we do not need to implement any small sample correction measures. Firstly, we examine the loss differential plots of the different combinations for all forecast horizons to verify covariance stationarity. As seen in Figure 15 below, which shows the loss differentials of the AR(4) and elastic net models, there is no observable trend or seasonality in any of the plots, other than the outliers during the 2008 Global Financial Crisis. Since the loss differential plots for all the other pairwise combinations look close to identical to the figure below, this suggests that all loss differentials under squared loss can be assumed to be covariance stationary to proceed with the DM test.

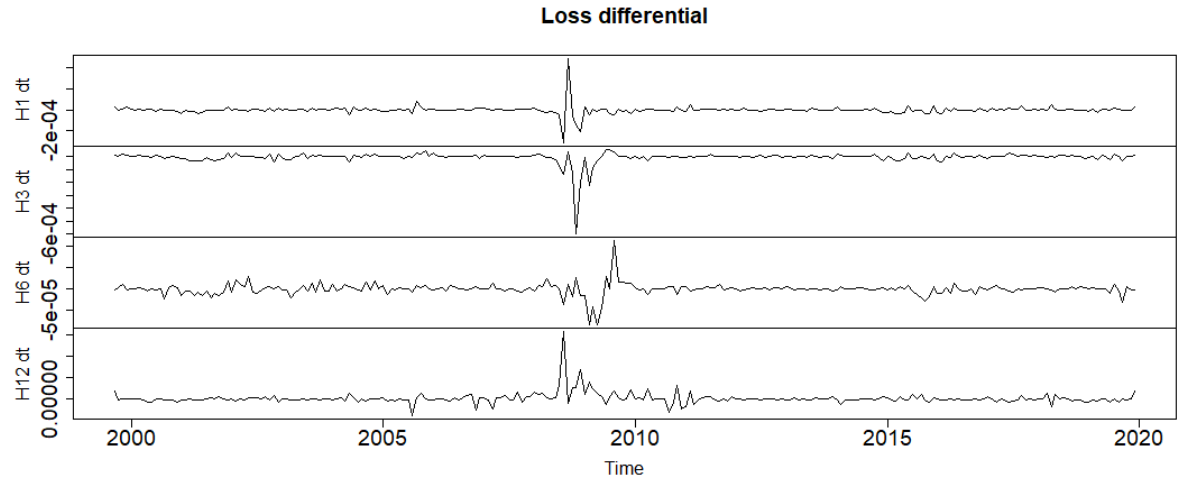

Fig 15: Loss differential plot: AR-Elastic Net

The following t-statistics were obtained from the DM equal predictive ability test. Overall, most of the t-statistics were not significant at the 5% significance level since their absolute

values were less than 1.96, and we cannot reject the null hypothesis, suggesting that they do not have different predictive abilities. For the 6-step ahead forecasts, the t-statistics of the AR-PLASSO and ElasticNet-PLASSO are negative and significant at the 5% significance level. For our RF tests, both tests against AR(4) and PLASSO are significant for all horizons and this is mainly due to the extremely low RMSE obtained from Random Forest.

Forecast Horizon	T-Statistic				
	AR - PLASSO(BIC)	AR - ElNet(BIC)	ElNet - PLASSO(BIC)	AR - RF	AR - Bagging
1-Step	-0.1575073	-0.970456	0.5085878	3.472907*	3.342414*
3-Step	-1.437976	-1.353514	0.947826	3.697444*	3.569156*
6-Step	-2.467984*	-1.051221	-2.141766*	3.142528*	3.089381*
12-Step	0.9530218	1.634116	0.4241291	2.983458*	2.932442*

Fig 16: DM Test t-statistics

5 Limitations

1. Missing Data:

Many variables related to consumption, housing, and manufacturing (as shown in the EDA) were omitted due to incomplete data. A more comprehensive dataset or imputation techniques might include these potentially valuable predictors, improving RMSE.

2. Computational constraints

Preparing large time series datasets for machine learning, including transformations can be computationally expensive. Computational costs and time also increase with dimensionality data as each predictor contributes to the complexity of the model.

6 Summary

In this project, we evaluated various machine learning (ML) methods and benchmark methods to forecast US GDP, using AR(4) as the baseline model. Among the models, Tree Based methods, such as Random Forest and Bagging, consistently outperforms the

benchmark and other ML models due to their ability to deal with highly correlated predictors and non-linear relationships between predictors and outcome variables.

LASSO and elastic net models performed poorly in comparison, likely due to the high density of predictive signals in our dataset. Both methods rely on regularisation, which can shrink important variables, particularly where there are many influential predictors. Additionally, we observed that the AR(4) model performed reasonably well for certain horizons such as 3-step and 6-step, indicating that linear models can capture GDP's short term dynamics, especially given the high relevance of INDPRO as a predictor in shorter horizons.

Forecast Horizon	AR(4) RMSE	Best Model	Best Model RMSE	DM-Test	Outperforms AR(4)?
1-step	0.6199	**LASSO**	**0.6194**	X	No
3-step	**0.6250**	**Post-LASSO**	0.6480	X	No
6-step	**0.6656**	**Elastic Net**	0.6754	X	No
12-step	0.6776	**Post-LASSO**	**0.6554**	X	No

Fig 17: Best models compared to AR(4)

*Random Forest and Bagging have been excluded due them clearly outperforming AR(4)

The table in figure 17 suggests that besides Random Forest and Bagging, which have significantly lower RMSE values than the rest of the models, the ML models generally are unable to outperform the benchmark model. While LASSO 1-step ahead and Post-LASSO 12-step ahead forecasts have a lower RMSE than AR(4), they were shown to be not statistically different from AR(4) through the DM-test.

As mentioned above, a surprising finding is that the Random Forest and Bagging have RMSE values that are 3 times smaller than the other ML models. Other interesting observations include the AR(4) model outperforming LASSO at the 6-step ahead forecast, and the Elastic Net model outperforming the PLASSO model specifically at the 6 steps ahead forecast, according to the DM-test.

In the field of machine learning, the success of predictive models often depends on specific conditions of the system and length of forecast horizon. From a study conducted by Chen, McCracken, and Ng (2022), they found that "when forecasting with a large number of predictors with mixed predictive power, density-based ML methods (such as bagging,

boosting, or neural networks) can somewhat outperform sparsity-based methods (such as Lasso) for short-horizon forecasts." However, they note that "it is not easy to distinguish the performance of these two types of methods for long-horizon forecasts." This aligns with the results of our projects as it suggests that while density-based methods like bagging or boosting may provide an edge in handling complex predictor interactions over shorter horizons, this advantage fades as the forecast horizon lengthens, highlighting the need to tailor model choice to the forecast objective.

7 Further Extensions

1. Dimension Reduction:

Exploring Principal Component Analysis (PCR) and Partial Least Squares (PLS) could help reduce dimensionality while retaining essential information, which also decreases the computational load without losing essential information. This will be helpful since it was proven that our indicators are strongly correlated.

2. Incorporating Deep Neural Networks

DNNs could be beneficial given the dense dataset, as they can handle numerous predictors by optimising weights across multiple layers, assigning higher weights to more predictive variables. Chu and Qureshi found that DNNs performed comparably to other complex models even with large predictor sets, due to their architecture's ability to balance the significance of each input (Chu, B., & Qureshi, S.,2022).

3. Mixed frequency Data Integration (MIDAS):

Using higher-frequency predictors, like weekly or daily data, could enhance the model's ability to capture short-term dynamics, especially for immediate GDP forecasts. Especially for the FRED-MD dataset, the monthly and quarterly indicators can limit the model's responsiveness to sudden shifts in the economic landscape. Hence, for forecasting US GDP, MIDAS models can combine macro-level, low-frequency indicators with more granular, high-frequency data.

7 Citations

1. Chu, B., & Qureshi, S. (2022). Comparing Out-of-Sample Performance of Machine Learning Methods to Forecast U.S. GDP Growth. *Computational Economics, 62*(1), 1567-1609. https://doi.org/10.1007/s10614-022-10312-z

2. Chen, Y., McCracken, M. W., & Ng, E. M. (2022). When forecasting with a large number of predictors with mixed predictive power, density-based ML methods (such as bagging, boosting, or neural networks) can somewhat outperform sparsity-based methods (such as Lasso) for short-horizon forecasts. *Computational Economics*. https://doi.org/10.1007/s10614-022-10312-z
3. Belloni, A., & Chernozhukov, V. (2013). It undoes the shrinkage bias by applying ordinary least squares (OLS) estimations to the predictors selected by LASSO. *Econometrica, 81*(4), 2231-2271. https://doi.org/10.3982/ECTA9626
4. Zhang, X., Smith, J., & Liu, Y. (2023). Emphasizing the importance of benchmarking forecasting models across diverse prediction horizons. *Journal of Forecasting*. https://doi.org/10.1002/for.2901