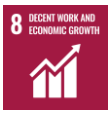

GDP GROWTH FORECASTING USING RANDOM FOREST AND XGBOOST MACHINE LEARNING ALGORITHMS

Research Paper
Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Public Policy

By: Arindam Misra
INDIA
(MEY200013)
Advisor: Prof. Fumio Hayashi

September 2021

Young Leaders Program (School of Government)
National Graduate Institute for Policy Studies
Tokyo, Japan

ACKNOWLEDGEMENT

I am grateful to Prof. Fumio Hayashi for being my advisor for this study. His valuable comments, suggestions, guidance and encouragement helped me to undertake and complete the study. His guidance and support for gathering the public data from the archives of Federal Reserve Bank of St. Louis enabled me to base this study entirely on public data.

I am also thankful to Prof. Kiyotaka Yokomichi for his constant help, support and comments during the multiple tutorials taken by him. The comments and suggestions have immensely helped to improve various aspects of the study and the paper.

I thank Prof. Mikitaka Masuyama for various sessions regarding the study and his advice to get out of comfort zone to try something new.

I also thank Ms. Yuko Ikenoya for her support throughout the course and this study.

I am also grateful to my family and especially my wife Rimjhim, whom I am lucky to have as a classmate and colleague as well. This study would have been impossible without her unconditional support and encouragement.

ABSTRACT

This paper applies Random Forest and Extreme Gradient Boosting (XGBoost) machine learning algorithms to forecast the Quarter over Quarter (QoQ) real Gross Domestic Product (GDP) growth in the US from Q1 2010 to Q1 2021. Vintage data from the archives of the Federal Reserve Bank of St. Louis, available on or before the forecast date have been used. The forecasts are made on the last Friday of the second and third months of the quarter and compared to the benchmark forecasts made by the New York Federal Reserve (NY Fed) on the same date. The study finds that for the period from Q1 2010 to Q1 2021 the forecasts made by the New York Federal Reserve are more accurate as compared to the predictions of machine learning algorithms. However, the accuracy of forecast for the pre-pandemic period (Q1 2010 to Q4 2019) is better for the Random Forest algorithm as compared to the forecast made by NY Fed. The Random Forest algorithm outperforms the XGBoost algorithm in prediction accuracy. This study suggests that machine learning methods can be used for forecasting important macroeconomic variables like real GDP growth.

1 INTRODUCTION

Knowledge about the current state of the economy is extremely important for policy makers. It enables them to take the required policy measures in a timely manner. Although many macroeconomic indicators are important for gauging the health of the economy, real GDP growth is one of the most important variables. Goal 8 of the Sustainable Development Goals aims to "Promote inclusive and sustainable economic growth, employment and decent work for all". In order to promote economic growth, it is imperative to measure or estimate it accurately, as indicated by the growth in real GDP. However, the official data for real GDP growth for a quarter in the US is available only after at least a month after the end of the quarter. Till then, the policy makers have to rely on the forecasts for the real GDP growth. This makes it important to have more accurate forecasts, so that a clearer and more reliable foresight is available while making important policy decisions for the economy.

Advances in the field of Econometrics and Statistics have contributed to various models that are used to make predictions about the GDP over varied periods of time. However, these models are conventional Econometrics based models which begin with an idea of reality that is to be modelled, and the parameters are calibrated based on theory, to make the predictions about GDP growth.

Machine Learning refers to a set of algorithms that improve automatically as they are fed with data. Instead of establishing a data-theory connection, as is the case with conventional prediction algorithms, machine learning is a purely empirical exercise which is based on 'data first' approach. Many economic variables change the way they interact with each other due to changes in technologies and other modifications over time. The conventional models which have an inherent data-theory connection might not be able to significantly capture and incorporate these

changes. Machine Learning on the other hand considers possibility of various ways in which even relatively small numbers of variables interact to give significant insights about the data. However, machine learning algorithms are more suitable for prediction problems like GDP forecasting rather than parameter estimation problems like relationship between unemployment and GDP growth. In Econometrics parlance, machine learning is more suitable for $\hat{y}$ problems rather than $\hat{\beta}$ problems (Mullainathan & Spiess, 2017).

This study uses machine learning algorithms to predict the QoQ growth of real GDP of the US. The XGBoost and Random Forest algorithms have been used in this study. These algorithms have been very successful in various Kaggle competitions and have been applied to prediction problems in various domains like predicting oil and wind energy prices. These algorithms also prevent overfitting through regularization and hyperparameters that control the depth of trees, which results in superior out-of-sample prediction performance (Bentéjac et al., 2019). The XGBoost algorithm has various algorithmic enhancements over traditional decision tree-based methods like tree pruning and regularization to prevent overfitting. Random Forest algorithm is better than conventional bagging techniques as it chooses a random subset of features for splitting the trees at each node. This leads to lower correlation amongst trees and improves the performance. These improvements add to the predictive power of the tree-based algorithms and make them the preferred choice for this study.

The study uses publicly available data for 15 variables from Archival FRED (ALFRED) which is a public data source of vintages maintained by the Federal Reserve Bank of St. Louis, and is used as input for the XGBoost and Random Forest machine learning algorithms after appropriate transformations. The study compares the forecasts made on the last Friday of the second and third months of the quarter to the benchmark forecasts made by the NY Fed on the same date, these specific dates in the quarter are called vantage points. As various economic variables which are used as inputs undergo revisions, and for fair comparison between the forecasts, only the vintages available on or before the date of prediction have been used. The study converts real GDP growth forecasting into a supervised learning problem. The data from Q3 2001 to one quarter before the quarter for which forecast is made, is used to train the Random Forest and XGBoost regressor with the QoQ GDP growth as the dependent variable.

The results show that the prediction accuracy of the NY Fed forecast for the period from Q1 2010 to Q1 2021 is higher than the XGBoost and Random Forest machine learning algorithms for both the vantage points, mainly due to the better forecasts during the pandemic period. The comparison between the forecasts made by the NY Fed and Random Forest algorithms during the period before the pandemic (Q1 2010 to Q4 2019) shows that Random Forest gives more accurate forecasts than NY Fed. Also, the XGBoost algorithm gives lower prediction accuracy than the Random Forest algorithm.

The machine learning algorithms fail to predict the phenomenal drop in the GDP during the pandemic. This can be attributed the inability of the tree-based machine learning regressors to extrapolate beyond the training data (Martius & Lampert, 2016) with the tree structure resulting in an upper and lower bound on the range of the regression function.

The remainder of this paper is as follows: Section 2 reviews the relevant literature in this area. Section 3 presents the methodology. Results are presented in Section 4 and Section 5 concludes the paper.

2 LITERATURE REVIEW

A large number of studies aimed at forecasting GDP have adopted the conventional approach which is based on techniques like time series econometrics, filtering and Big Data analytics. The conventional methods of forecasting are based on stochastic models which make various assumptions regarding the distribution of the independent variables and the relationship between them to predict the dependent variable. These models try to establish a data-theory connection and can explain the impact of individual variables on the dependent variable[1]. However, they might not capture the interactions between variables unless specific interaction terms are included in the model (Mullainathan & Spiess, 2017).

Application of machine learning to prediction of GDP is a relatively new approach and many prior studies have indicated significant enhancement in the prediction accuracy as compared to the conventional approach. There have also been studies that have reported deterioration in prediction accuracy for some machine learning methods as compared to the traditional approaches (Jung et al., 2018).

Various prior studies have used machine learning to forecast US GDP growth. Soybilgen and Yazgan (2021) have used bagged decision trees, random forests, and stochastic gradient tree boosting after extracting factors from 10 variable groups using a dynamic factor model. They use FRED-MD data set[2] to nowcast US GDP between 2000Q2 and 2018Q4 and compare the predictions with GDPNow – the nowcasting model of Atlanta Fed, and report improved performance as compared to the benchmark forecast. However, their study does not give results of machine learning algorithms for the pandemic period. Loermann and Maas (2019) have also forecasted the US GDP using feed forward Artificial Neural Networks and the FRED-MD database. They compare the results to the state-of-the-art dynamic factor models and Survey of Professional Forecasters and report an improvement in forecast performance. Nyman and Ormerod (2016) have used the Random Forest algorithm to forecast the one, three and six quarters ahead US GDP growth during 1990Q2 to 2016Q2 and tried to predict the recession of 2009. They found that

[1] See Banbura et al. (2013) for a detailed exposition of the conventional approach and an application to GDP in the Euro area. Bok et al. (2018) provides a more recent review.

[2] FRED-MD is a large macroeconomic database of monthly variables by St. Louis Fed for empirical studies.

recession in the first half of 2009 could have been predicted six quarters earlier in 2007. Rajkumar (2017) compare the performance of OLS, Logistic Regression, Random Forest and Neural Networks in predicting the surprises to US GDP growth. They reported that Neural Networks perform the best in forecasting steep rise or fall in GDP growth. Wochner (2020) has used a theory-led machine learning framework by synthesizing machine learning algorithms, dynamic factor model and business cycle literature using the FRED-MD database and reported improvements in forecast accuracy in expansionary as well as recessionary periods.

Machine Learning has also been used to forecast GDP growth of many other countries. Richardson et al. (2018) attempted to nowcast New Zealand GDP based on Machine Learning algorithms and found improvement in the predictions as compared to the conventional approach. Jung et al. (2018) tried Machine Learning algorithms like Elastic Net, Recurrent Neural Networks and Super Learner algorithms out of which super learner algorithm, which is a combination of various machine learning methods performed better than the conventional techniques of prediction. Kurihara and Fukushima (2019) applied machine learning to forecast GDP and CPI for G7 countries. Woloszko (2020) used adaptive tree-based approach for GDP forecasting. Qureshi et al. (2020) have reported improvements in forecasts using XGBoost algorithm to predict the Canadian GDP using official and Google Trends data. Yoon (2020) has applied XGBoost and Random Forest algorithms to forecast Japanese GDP and reported improved performance.

This study contributes to the existing literature as it uses the archival data from ALFRED to base the forecasts on vintages released on or before the forecast date instead of using the FRED-MD database. This helps in accurately creating the quarterly variables without any missing values due to reporting lag or any other reasons. This also ensures fair comparison with benchmark forecasts. This study also uses variables related to the real economy after transforms to make them stationary, instead of the factor variables and principal components used by Soybilgen and Yazgan (2021) and Qureshi et al. (2020) respectively. This study also evaluates the performance of Random Forest and XGBoost algorithms during the turbulent period of the pandemic as well relatively stable macroeconomic environment. Also, this study compares the machine learning forecasts with the benchmark forecast of the New York Federal Reserve unlike other benchmark forecasts used in other studies.

3 METHODOLOGY

This study uses the Random Forest and XGBoost machine learning algorithms to make QoQ real GDP growth forecasts for the US economy from Q1 2010 to Q1 2021. In order to apply machine learning to the problem of GDP forecasting, the problem of forecasting needs to be converted into a supervised learning problem. In supervised learning, the independent variables are used as input for the machine learning algorithm and the dependent variable is used as the label or outcome. The algorithm gets trained using the labels and it builds regression trees which predict

the dependent variable when independent variables are given as inputs for prediction. In this study the various independent variables which have an impact on GDP growth are used as input to the XGBoost and Random Forest algorithms and the QoQ real GDP growth (dependent variable) is assigned as the label or outcome to train the algorithm. In order to make prediction for QoQ real GDP growth for quarter n, the algorithm is trained from Q3 2001 to Q_{n-1}. The data for Quarter n is then used to make prediction for that quarter using the model which has been trained using the data till the previous quarter.

The forecasts are made on the last Friday of the second and third month of the quarter. For Example, for the 3rd quarter of 2019 which is from 1st July 2019 – 30th September 2019 (the first official GDP growth figures for which were released on 30th October 2019) the forecasts are made in the second and third month of the quarter on 30th August 2019 and 27th September 2019 respectively (vantage points). Similarly, the vantage points for 4th Quarter of 2019 are 29th November 2019 and 27th December 2019. These vantage points were chosen as many of the monthly input variables required for forecasting the current quarter GDP growth are available by these dates and it also coincides with the forecast dates of the NY Fed.

In order to make a fair comparison with the forecasts of the NY Fed, it is imperative that only the data for the monthly independent variables available on or before the forecast date is used. For this purpose, this study uses the archival data from ALFRED released on specific dates, called vintages. This study makes sure that no vintages beyond the prediction date are used for forecasting. If due to some reporting delays or some other reasons the vintages are not available in time, the earlier available vintages are used. The predictions made by the machine learning algorithms are compared to the predictions of the NY Fed on the same date and the first official release of the QoQ real GDP growth vintage for that quarter.

3.1 DATA

For forecasting the QoQ real GDP growth, various vintages for the 15 variables were collected from ALFRED from Q1 2001 to Q1 2021. The study uses 14 monthly variables and QoQ real GDP growth as the only quarterly variable for forecasting. The variables include indicators of real economic activity as well as economic outlook surveys. The list of variables used includes variables prominently featured in Bok et al. (2018) and standard transforms available in nowcasting literature have been applied. Table 1 gives the list of variables, the lags with which each variable is published and the transformations applied to each variable. All the variables in Table 1 except QoQ real GDP growth have a monthly frequency and are seasonally adjusted. In order to make the variables stationary, the first differences of the variables have been taken in log or levels. Logistic transformations have been applied for the Empire State Manufacturing Survey and Manufacturing Business Outlook Survey but no first differences have been taken as the survey is about changes. The lag of QoQ real GDP growth is also used as a regressor for predicting the QoQ real GDP growth in the current quarter.

Table 1
Variables Used and Transformations

S. No	Name	Unit	ALFRED Code	Lag	Transform
1	QoQ real GDP growth	Percentage change	A191RL1Q225SBEA	30 days	
2	Industrial Production Index	Index	INDPRO	15 days	$\Delta \log$
3	Real Disposable Personal Income	Constant USD	DSPIC96	30 days	$\Delta \log$
4	Building Permits	Units	PERMIT	15 days	$\Delta \log$
5	Housing Starts	Units	HOUST	15 days	$\Delta \log$
6	Business Inventories	Current USD	BUSINV	45 days	$\Delta \log$
7	Durable Goods Orders	Current USD	DGORDER	30 days	$\Delta \log$
8	New One Family Houses Sold	Units	HSN1F	30 days	$\Delta \log$
9	Advance Retail Sales	Current USD	RSAFS	15 days	$\Delta \log$
10	Real Personal Consumption Expenditures	Constant USD	PCEC96	30 days	$\Delta \log$
11	Total Nonfarm Private Payroll Employment	Number of persons	NPPTTL	3-4 days	$\Delta \log$
12	Unemployment Rate	Percentage	UNRATE	7 days	Δ
13	Capacity Utilization	Percentage	TCU	15 days	$100 \times log\left[\frac{100 - x}{x}\right]$
14	Empire State Manufacturing Survey	Diffusion Index	GACDISA066MSFRBNY	-14 days	$100 \times log\left[\frac{100 - x}{x + 100}\right]$
15	Manufacturing Business Outlook Survey	Diffusion Index	GACDFSA066MSFRBPHI	-11 days	$100 \times log\left[\frac{100 - x}{x + 100}\right]$

The vintages of the variables given in Table 1 are collected for the two different vantage points in the following manner:

For the vantage point on the last Friday of the third month of the quarter, the monthly variables are transformed into quarterly variables. The average of the three most recently available monthly observations on or before the forecast date is taken as the quarterly value of that variable. For Example, for the 3rd quarter of 2019 for which the forecast date of the vantage point on the last Friday of September (third month) is 27th September 2019, the vintage for September 2019 for Advance Retail Sales is not available on the date of forecast. Thus, the average of the values for June, July and August (released on 13th September 2019) are taken before first differencing, as the quarterly variable for predicting the Q3 2019 QoQ real GDP growth. The other quarterly variables are created in the same manner.

For the vantage point on the last Friday of the second month of the quarter the latest available value for the variables 2-13 of Table 1 on or before the date of the forecast were used for predicting the QoQ real GDP growth. For Example, for the 3rd quarter of 2019 for which the forecast date of the vantage point on the last Friday of August (second month) is 30th August 2019, the vintage for August 2019 for Advance Retail Sales is not available on the date of forecast. Thus, the available monthly value for July (vintage released on 15th August 2019) is used for forecast. For the Empire State Manufacturing Survey and Manufacturing Business Outlook Survey the values for the first month and the second month of the quarter are available before the date of the forecast, and thus, the average of the two months is taken for the variables and then the logistic transformation given in Table 1 is applied for the forecast.

3.2 THEORETICAL FRAMEWORK AND APPLICATION

This section summarizes the XGBoost and Random Forest algorithms and their theoretical background. A more detailed discussion on XGBoost and Random Forest algorithms can be found in the original papers by Chen and Guestrin (2016), Breiman (2001) and Hastie et al. (2013).

The XGBoost and Random Forest algorithms are regression tree-based algorithms. The prediction function is like a tree and it splits into two sub-trees or leaves at each node. This split is based on the value of an independent variable, which determines whether the left or the right tree is to be considered. Each leaf node has a prediction associated with it. Figure 1 shows a regression tree structure and how predictions are made using the tree. This property of trees enables creation of multiple interactions between variables, which is a unique ability of machine learning algorithms and increases the dimensionality of the prediction problem substantially.

Using the above approach, it might be possible to get a perfectly fitted tree which would have a leaf for every observation in the training data. However, this gives rise to a problem in machine learning which is called overfitting. The choice of the

function which gives the best in-sample fit can give very poor out-of-sample performance. This happens as the function is no longer general and overfits the

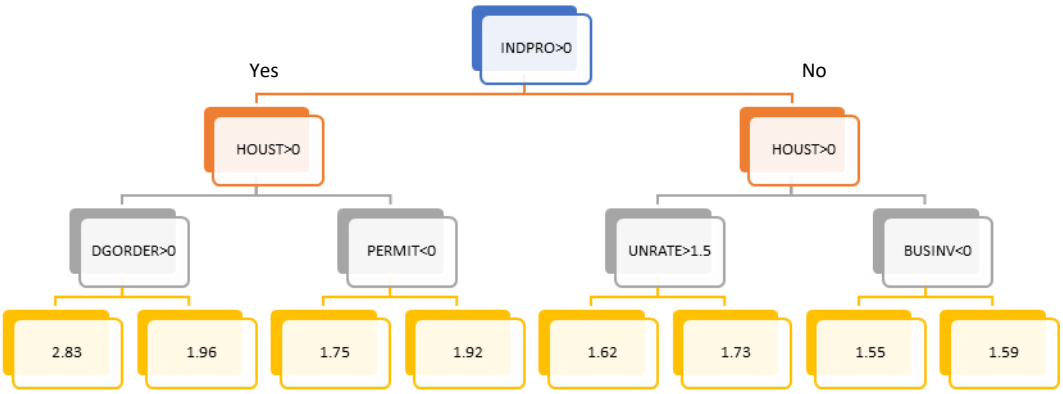

Figure 1: Prediction using a regression tree

training data. In order to overcome this problem, the algorithms have a regularization term in their cost function, which controls the overall complexity of the model.

Another important aspect of predictions using machine learning methods is cross validation and hyperparameter tuning. Hyperparameters are the various parameters of the machine learning algorithms that can be specified by the user. Cross Validation is a process in which the training data set is divided into various blocks of test and train data sets which are known as 'folds'. The performance of the machine learning algorithm is averaged over all the folds or iterations and it gives an indication of the out-of-sample prediction accuracy of the algorithm for a given set of hyperparameters. In many machine learning algorithms k-fold cross validation is used in which the data is divided into k parts and over multiple iterations each of the k parts is used for testing, while the remaining parts are used for training.

This cross-validation technique is not suitable for time series problems like GDP forecasting as there are temporal dependencies in data, and future data should not be used to test past data. Thus, this study uses the method of walk forward validation. Figure 2 shows the difference between 4-fold cross validation and a walk forward validation with 4 folds.

In this study the XGBoost and Random Forest algorithms have been implemented in Python 3.8 using the Scikit-learn machine learning library version 0.24.1[3].

[3] The data and code are available on Github at https://github.com/arindammisra/ML-GDP

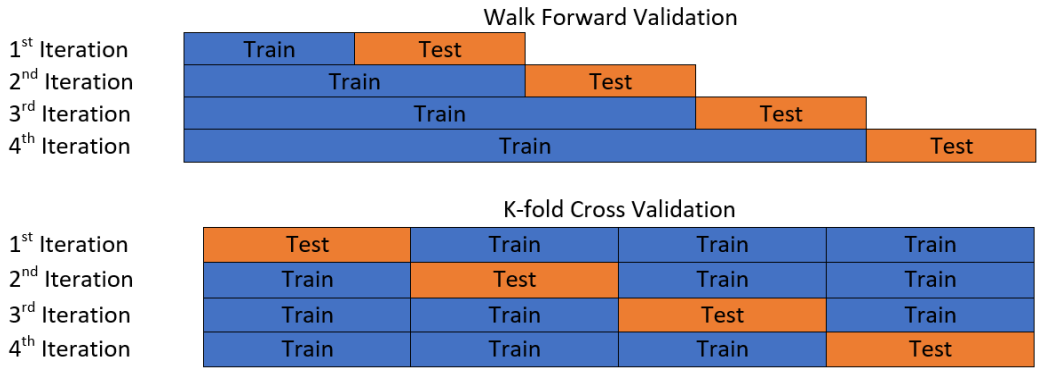

Figure 2

3.2.1 XGBOOST ALGORITHM

XGBoost is an implementation of optimized gradient boosted trees which takes training set $\{(x_i, y_i)\}_{i=1}^{n}$ as input. It is an ensemble technique which uses a number of gradient boosted trees to improve the predictive performance.

XGBoost makes various improvements over conventional regression trees. It makes multiple trees depending on the previous iterations, either till the number of iterations specified as a hyperparameter are reached or there are no improvements in the training by making more trees. XGBoost also has tree pruning feature a regularization function as a part of its objective function which takes into account the complexity of the tree (Chen & Guestrin, 2016). If $\hat{y}_i^{(t)}$ is the prediction of the i^{th} instance at the t^{th} iteration, then XGBoost tries to find f_t that minimizes the following objective function.

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

The first term measures how well the model fits the training data and the second term measures the complexity of the trees. XGBoost uses two parameters to define the complexity of the tree: i) The number of leaves and ii) L2 norm of leaf scores. In the expression below, λ and γ are hyperparameters, K is the number of leaves and w_j is the weight assigned to the leaf.

$$\Omega(f_t) = \gamma K + \frac{1}{2}\lambda \sum_{j=1}^{K} w_j^2$$

The steps in gradient boosting as given by Hastie et al. (2013) are given below:

Where, $L(y_i, f(x))$ is the differentiable loss function, r_{im} is the pseudo residual for the i^{th} observation in the m^{th} iteration and γ is a predicted value.

Step 1: Initialize $f_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$
Step 2: For m=1 to M:
a) For i=1, 2,…,N compute pseudo residual $r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$
b) Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm}, j = 1,2, \dots, J_m$
c) For $j = 1,2, \dots, J_m$ compute $\gamma_{jm} = argmin_\gamma \sum_{x_{i \in R_{jm}}} L(y_i, f_{m-1}(x_i) + \gamma)$
d) Update $f_m(x) = f_{m-1}(x) + \eta \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$
Step 3: Output $\hat{f}(x) = f_m(x)$

The XGBoost algorithm uses a parameter η which prevents overfitting and determines the rate at which learning takes place. However, due to its tree-based structure XGBoost is unable to perform extrapolation.

3.2.2 RANDOM FOREST ALGORITHM

Random Forest is a very popular ensemble machine learning algorithm which uses Bootstarp Aggregation or bagging instead of boosting. It is also based on regression trees. However, Random Forest algorithm builds multiple regression trees, trains them independently and gives the average of the predictions from the constituent trees (Breiman, 2001). A random subset of features is chosen to split the tree at each node, resulting in lower correlation between trees. This leads to improvement in prediction performance. However, due to its tree-based structure, Random Forest is also not able to extrapolate. The basic steps of the Random Forest algorithm which takes training set $\{(x_i, y_i)\}_{i=1}^{n}$ as input are given below:

Step 1: Select j out of the i features (j<i) randomly for the set x_i
Step 2: Find the best split point p and variable among the j variables
Step 3: Split the node into daughter nodes using the best split that minimizes the mean square error $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$.
Step 4: Repeat the steps 1 to 3 till the minimum node size is reached (The minimum node size is the minimum number of elements at a node after which no splits are performed)
Step 5: Repeat the steps 1to 4 to create n trees for building the forest
Step 6: Calculate the final output by averaging the output of all the trees in the ensemble.

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} F_i(x)$$

3.2.3 HYPERPARAMETER TUNING

The algorithms have a number of hyperparameters which can be adjusted by the users for improving performance. The learning rate parameter shrinks the feature weights after each boosting iteration to make the gradient boosting more conservative. maximum depth parameter determines how deep the trees can get. The number of estimators parameter determines the number of trees that the

algorithm uses. These parameters need to be tuned to get best performing forecast model for each prediction. GridsearchCV function of the Scikit Learn library has been used in this study to tune the parameters for the XGBoost as well as Random Forest models. The function performs a grid search over all the parameter sets which are given and all the combinations of hyperparameters are tried and the model with best hyperparameters is used for prediction. The hyperparameter values which are used for grid search are given in Table 2.

Table 2
Hyperparameter Sets

Algorithm	Hyperparameter	Value Set
XGBoost	Number of estimators	100, 500, 1000
	Maximum tree depth	1, 2, 3, 4, 5, 6, 7, 8, 9
	Learning Rate	0.01, 0.1, 0.3
Random Forest	Number of estimators	100, 500, 1000
	Maximum tree depth	1, 2, 3, 4, 5, 6, 7, 8, 9

4 RESULTS

The forecasts made by the XGBoost and Random Forest algorithms as well as the forecasts of the NY Fed and the first release of official QoQ real GDP growth figure for both the vantages along with the dates of forecast are given in Appendix A.The comparison between forecasts is done using the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). They are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \qquad RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

y_i is the actual QoQ real GDP growth for the quarter and $\hat{y}_i$ is the forecast. N is the total number of observations. The results show that for the period from Q1 2010 to Q1 2021 the forecasts made by the NY Fed are more accurate as its MAE and RMSE values are much lower than both the machine learning algorithms for both the vantage points, as shown in Table 3.

Table 3
Comparison between forecasts for the total period and pre-pandemic period

Total Period					
Second Month Vantage			Third Month Vantage		
	MAE	RMSE		MAE	RMSE
NY Fed	1.78	3.33	NY Fed	2.00	4.17
XGBoost	2.70	6.59	XGBoost	2.69	6.56
Random Forest	2.51	6.51	Random Forest	2.55	6.64

Pre-Pandemic					
Second Month Vantage			Third Month Vantage		
	MAE	RMSE		MAE	RMSE
NY Fed	1.23	1.72	NY Fed	1.14	1.58
XGBoost	1.25	1.71	XGBoost	1.22	1.64
Random Forest	1.03	1.47	Random Forest	1.01	1.42

This is mainly because of the large forecasting errors made by the Random Forest and XGBoost machine learning algorithms in the pandemic period (Q1 2020 to Q1 2021). The machine learning algorithms fail to forecast the unprecedented dip and rise of ~30% over a span of two quarters. This can be explained by the difficulty in extrapolation faced by classical machine learning algorithms like XGBoost and Random Forest (Martius & Lampert, 2016). When the algorithms are presented with data which lies beyond the limits of their training data, the predictions are not accurate. This happens mainly because the tree-based algorithms use the independent variables to partition the data and assign an appropriate value to the dependent variable. This results in only certain finite values of the dependent variable being predicted, which do not form a continuous range of the prediction function in real space. This results in upper and lower bounds of the predicted values beyond a certain input level of the regressors. This makes the algorithms unable to accurately predict steep rise or falls in GDP growth and give results proportionate to the steep rise or fall in values of the regressors. Rajkumar (2017) found that Random Forest algorithm did not accurately predict the surprises to US GDP growth. Random Forest algorithm tries to fit a tree on the given data with robust splitting decision. With n regressors, the tree can be split in at most n ways at each node resulting in a limited set of forecasts at each leaf node. The study also reported that this shortcoming can be overcome using neural networks.

Amongst the two machine learning algorithms Random Forest algorithm performs better than XGBoost algorithm for period from Q1 2010 to Q1 2021. The graphs showing the comparison of forecasts for both the vantages for the period Q1 2010 to Q1 2021 are given in Figure 3 and 4. The second month vantage forecasts for the period Q1 2010 to Q1 2021 are more accurate than the forecasts for the third month vantage. The high difference in RMSE and MAE for the NY Fed forecast is mainly because of the forecast revision for Q2 2020 from -35.53% to -16.33% in the second month to the third month. The difference in forecasts during the pandemic period is also the reason for lower RMSE and MAE for the second month vantage than the third month vantage for the machine learning algorithms.

To analyse the performance in less turbulent times, comparison has also been made between the forecasts of NY Fed and machine learning algorithms during the pre-pandemic period (Q1 2010 to Q4 2019) for both the vantage points, which are also given in Table 3. It is found that the Random Forest algorithm gives better forecasts than the NY Fed for both the vantage points. The XGBoost algorithm does not perform well for both the vantage points in the pre-pandemic period as well and its

MAE and RMSE values are higher than the forecasts of NY Fed and Random Forest forecasts. The graphs showing the comparison of forecasts for both the vantages for the period Q1 2010 to Q4 2019 are given in Figure 5 and 6.

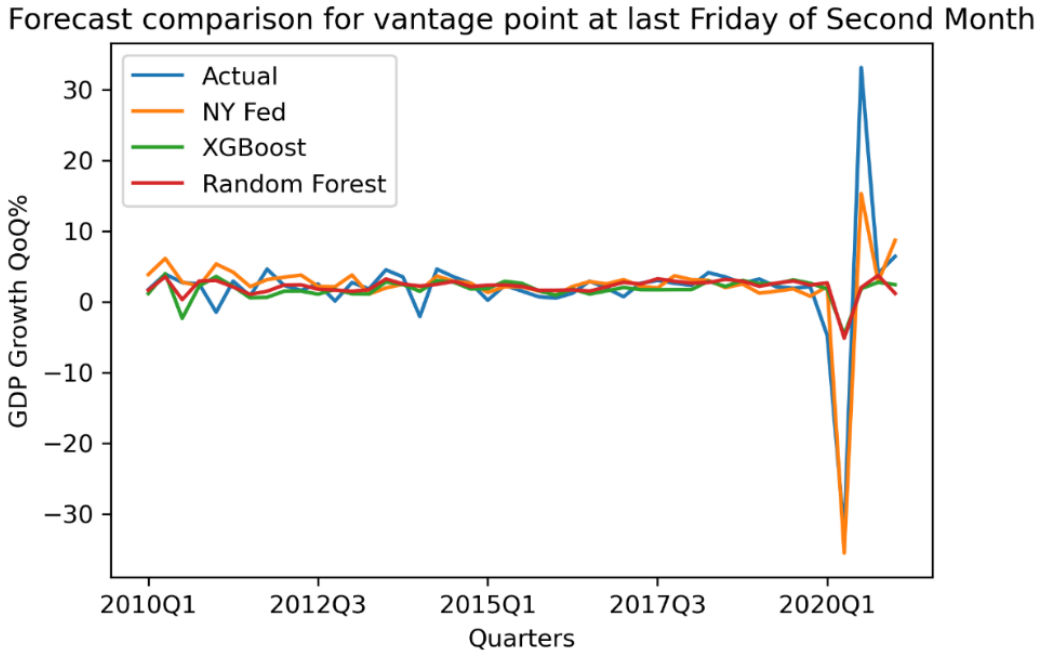

Figure 3

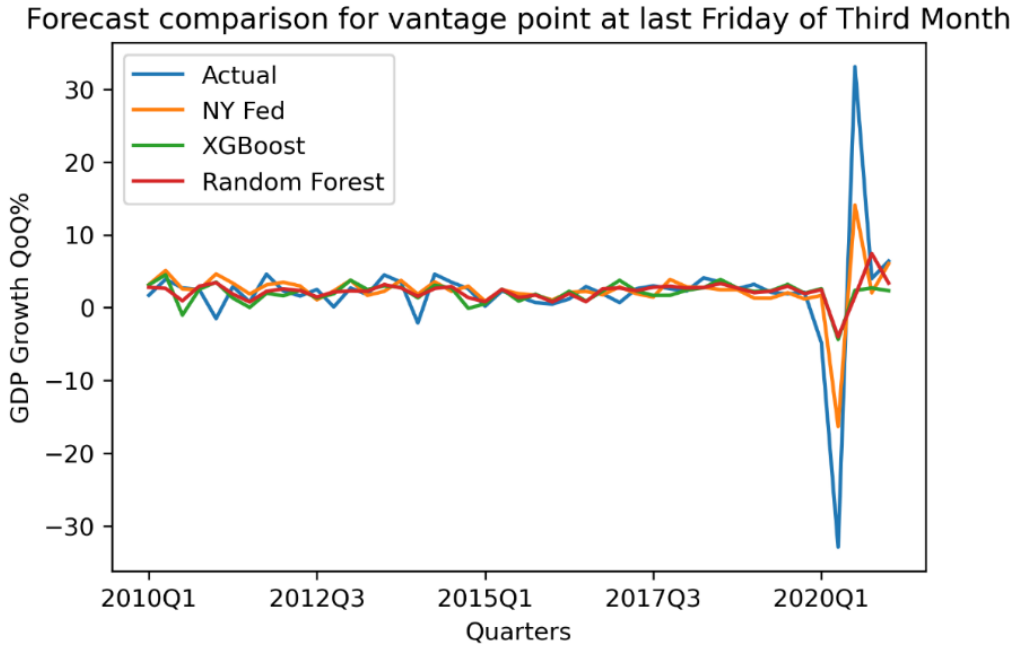

Figure 4

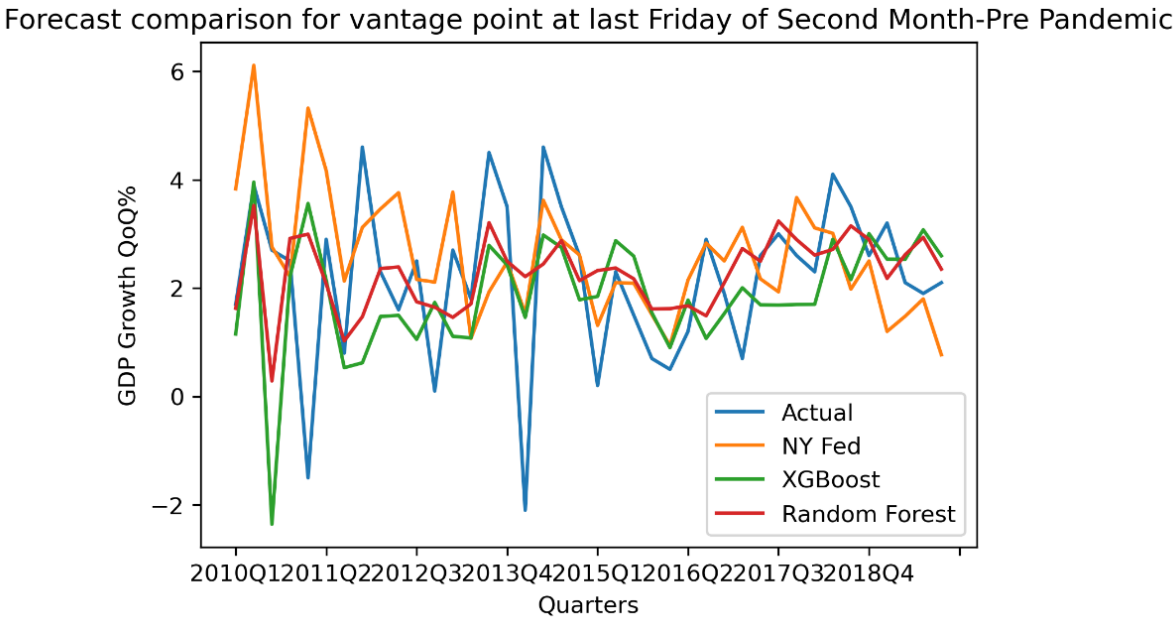

Figure 5

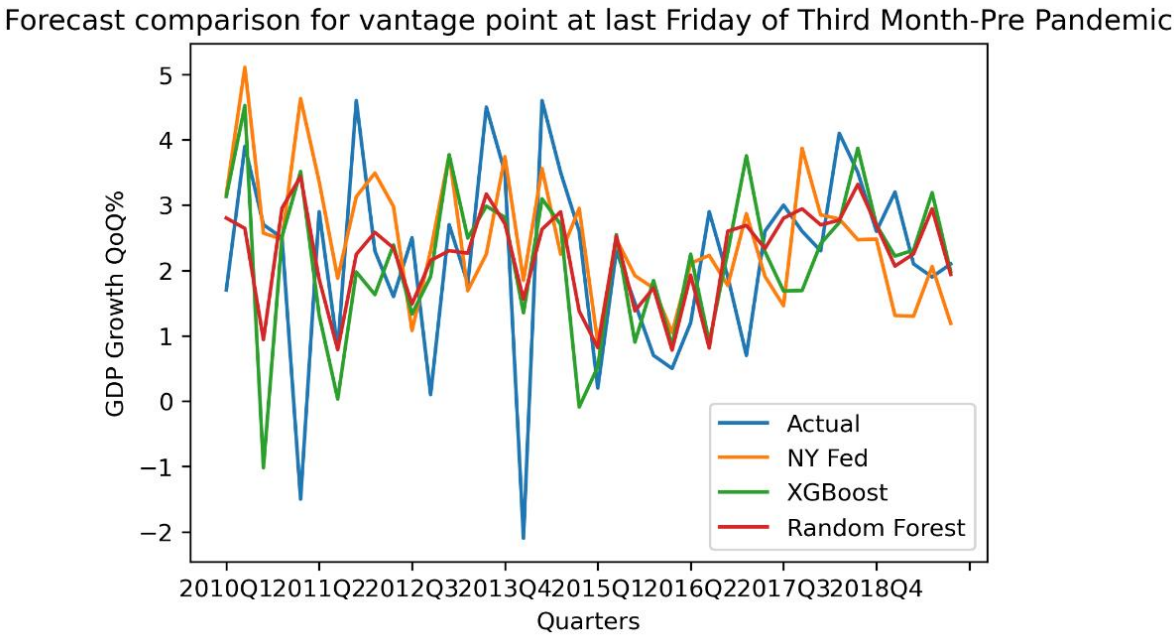

Figure 6

In the period before the pandemic, the third month vantage forecasts for the NY Fed and machine learning algorithms are more accurate then the second month vantage forecasts. This can be attributed to the more monthly data points used to

make the quarterly input variables which capture larger part of the quarter as compared to the second month vantage.

5 CONCLUSION

This study applies machine learning to the problem of real GDP growth forecasting. It finds that although the forecasting performance of machine learning algorithms for the total forecast period from Q1 2010 to Q1 2021, which includes the period of the Covid-19 pandemic that has had a significant economic impact, is not as good as the benchmark forecasts of the NY Fed for the same period. However, the Random Forest algorithm gives better out-of-sample forecast performance for the period excluding the pandemic. This suggests that machine learning algorithms can have higher prediction power for certain problems and they can be used for forecasting important macroeconomic variables like GDP growth.

Machine learning does not try to establish causal relationships between variables and it is very difficult to ascertain the impact and contribution of individual independent variables on the dependent variable. However, their power to predict better can be exploited to a greater extent by using them more often or creating ensemble models which use machine learning along with the conventional econometric models. Moreover, as machine learning models do not extrapolate well and fail to give accurate predictions for outlier events like the Covid-19 pandemic, some additional variables, training and further research would be required to improve the predictions during the turbulent times.

As this study tries to improve forecast accuracy using machine learning, it can enable better prediction of various important macroeconomic variables. This can help policymakers to ascertain the current health the economy before the official data is released and take timely policy actions. This also contributes to the 8th Sustainable Development Goal - “Promote inclusive and sustainable economic growth, employment and decent work for all” by giving more precise estimates of economic growth.

6 REFERENCES

Bańbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-casting and the real-time data flow. In *Handbook of economic forecasting* (Vol. 2, pp. 195-237). Elsevier.

Bentéjac, C., Csörgo, A., & Martínez-Muñoz, G. (2019). A Comparative Analysis of XGBoost. *ArXiv, abs/1911.01914*.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with Big Data. *Annual Review of Economics*, *10*(830), 615–643.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jung, J.K., Patnam, M., & Ter-Martirosyan, A. (2018). An algorithmic crystal ball: forecasts-based on machine learning. *IMF Working Papers*, *18*(230). https://doi.org/10.5089/9781484380635.001

Kurihara, Y., & Fukushima, A. (2019). AR model or machine learning for forecasting GDP and consumer price for G7 Countries. *Applied Economics and Finance*, *6*(3).

Loermann, J., & Maas, B. (2019), Nowcasting US GDP with artificial neural networks, *MPRA Paper 95459, University Library of Munich, Germany*,

Martius, G., & Lampert, C. H. (2016). Extrapolation and learning equations. *arXiv preprint arXiv:1610.02995*.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106. DOI: https://www.doi.org/10.1257/jep.31.2.87

Nyman, R., & Ormerod, P. (2017). Predicting economic recessions using machine learning algorithms. *arXiv preprint arXiv:1701.01428*.

Qureshi, S., Chu, B. M., & Demers, F. S. (2020). Forecasting Canadian GDP growth using XGBoost. *Carleton Economic Papers 20-14, Carleton University, Department of Economics*

Rajkumar, V. (2017). *Predicting surprises to GDP: a comparison of econometric and machine learning techniques* (Doctoral dissertation, Massachusetts Institute of Technology).

Richardson, A., Mulder, T.V.F., Vehbi.T. (2018). Nowcasting New Zealand GDP using machine learning algorithms. *Centre for Applied Macroeconomic Analysis, CAMA Working Paper 47/2018.*

Soybilgen, B., & Yazgan, E. (2021). Nowcasting US GDP Using Tree-Based Ensemble Models and Dynamic Factors. *Computational Economics*, *57*(1), 387-417.

Wochner, D. (2020). Dynamic factor trees and forests–a theory-led machine learning framework for non-linear and state-dependent short-term us gdp growth predictions. *KOF Working Papers, 472*.

Woloszko, N. (2020). Adaptive trees: a new approach to economic forecasting. *Organisation for Economic Co-operation and Development, Working papers no. 1593*, 1–43.

Yoon, J. (2021). Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, *57*(1), 247-265.

Appendix A

Table 3
Forecast comparisons and forecast dates for second month vantage point

Quarter	NY Fed Forecast Date	Actual	NY Fed	XGBoost	Random Forest
2010Q1	26-Feb-10	1.7	3.83	1.15	1.63
2010Q2	28-May-10	3.9	6.11	3.96	3.52
2010Q3	27-Aug-10	2.7	2.76	-2.36	0.29
2010Q4	26-Nov-10	2.5	2.21	2.21	2.92
2011Q1	25-Feb-11	-1.5	5.32	3.56	3.00
2011Q2	27-May-11	2.9	4.17	2.15	2.08
2011Q3	26-Aug-11	0.8	2.13	0.53	1.02
2011Q4	25-Nov-11	4.6	3.12	0.62	1.47
2012Q1	24-Feb-12	2.3	3.46	1.48	2.36
2012Q2	25-May-12	1.6	3.76	1.50	2.39
2012Q3	31-Aug-12	2.5	2.16	1.05	1.75
2012Q4	30-Nov-12	0.1	2.11	1.73	1.65
2013Q1	22-Feb-13	2.7	3.77	1.11	1.46
2013Q2	31-May-13	1.8	1.07	1.08	1.71
2013Q3	30-Aug-13	4.5	1.93	2.79	3.21
2013Q4	29-Nov-13	3.5	2.48	2.43	2.49
2014Q1	28-Feb-14	-2.1	1.53	1.46	2.21
2014Q2	30-May-14	4.6	3.62	2.98	2.44
2014Q3	29-Aug-14	3.5	2.89	2.75	2.88
2014Q4	28-Nov-14	2.6	2.61	1.78	2.14
2015Q1	27-Feb-15	0.2	1.31	1.85	2.32
2015Q2	29-May-15	2.3	2.10	2.88	2.37
2015Q3	28-Aug-15	1.5	2.09	2.59	2.17
2015Q4	27-Nov-15	0.7	1.51	1.54	1.61
2016Q1	26-Feb-16	0.5	0.94	0.90	1.62
2016Q2	27-May-16	1.2	2.15	1.78	1.67
2016Q3	26-Aug-16	2.9	2.83	1.07	1.49
2016Q4	25-Nov-16	1.9	2.50	1.53	2.10
2017Q1	24-Feb-17	0.7	3.12	2.01	2.73
2017Q2	26-May-17	2.6	2.17	1.69	2.49
2017Q3	25-Aug-17	3	1.93	1.69	3.24
2017Q4	24-Nov-17	2.6	3.67	1.70	2.89
2018Q1	23-Feb-18	2.3	3.11	1.70	2.61
2018Q2	25-May-18	4.1	3.01	2.90	2.71
2018Q3	31-Aug-18	3.5	1.98	2.16	3.15
2018Q4	30-Nov-18	2.6	2.50	3.00	2.90
2019Q1	22-Feb-19	3.2	1.20	2.53	2.18
2019Q2	31-May-19	2.1	1.48	2.53	2.62
2019Q3	30-Aug-19	1.9	1.80	3.08	2.94
2019Q4	29-Nov-19	2.1	0.77	2.60	2.35
2020Q1	28-Feb-20	-4.8	2.14	1.80	2.65

2020Q2	29-May-20	-32.9	-35.53	-4.69	-5.16
2020Q3	28-Aug-20	33.1	15.27	1.84	2.00
2020Q4	27-Nov-20	4	2.82	2.74	3.70
2021Q1	26-Feb-21	6.4	8.68	2.40	1.16

Table 4
Forecast comparisons and forecast dates for third month vantage point

Date	NY Fed Forecast Date	Actual	NY Fed	XGBoost	Random Forest
2010Q1	26-Mar-10	1.7	3.17	3.13	2.80
2010Q2	25-Jun-10	3.9	5.11	4.52	2.64
2010Q3	24-Sep-10	2.7	2.57	-1.02	0.94
2010Q4	31-Dec-10	2.5	2.48	2.52	2.95
2011Q1	25-Mar-11	-1.5	4.63	3.51	3.44
2011Q2	24-Jun-11	2.9	3.36	1.32	1.87
2011Q3	30-Sep-11	0.8	1.88	0.03	0.79
2011Q4	30-Dec-11	4.6	3.13	1.98	2.25
2012Q1	30-Mar-12	2.3	3.49	1.63	2.58
2012Q2	29-Jun-12	1.6	2.98	2.39	2.35
2012Q3	28-Sep-12	2.5	1.08	1.33	1.49
2012Q4	28-Dec-12	0.1	2.30	1.91	2.15
2013Q1	29-Mar-13	2.7	3.77	3.77	2.30
2013Q2	28-Jun-13	1.8	1.69	2.50	2.26
2013Q3	27-Sep-13	4.5	2.25	2.98	3.17
2013Q4	27-Dec-13	3.5	3.74	2.81	2.71
2014Q1	28-Mar-14	-2.1	1.85	1.35	1.56
2014Q2	27-Jun-14	4.6	3.56	3.10	2.63
2014Q3	26-Sep-14	3.5	2.25	2.70	2.90
2014Q4	26-Dec-14	2.6	2.95	-0.09	1.38
2015Q1	27-Mar-15	0.2	0.89	0.52	0.82
2015Q2	26-Jun-15	2.3	2.48	2.55	2.52
2015Q3	25-Sep-15	1.5	1.92	0.90	1.38
2015Q4	25-Dec-15	0.7	1.73	1.85	1.73
2016Q1	25-Mar-16	0.5	1.05	0.85	0.78
2016Q2	24-Jun-16	1.2	2.11	2.25	1.93
2016Q3	30-Sep-16	2.9	2.23	0.91	0.81
2016Q4	30-Dec-16	1.9	1.77	2.29	2.60
2017Q1	31-Mar-17	0.7	2.87	3.75	2.69
2017Q2	30-Jun-17	2.6	1.91	2.29	2.33
2017Q3	29-Sep-17	3	1.46	1.69	2.80
2017Q4	29-Dec-17	2.6	3.87	1.69	2.94
2018Q1	23-Mar-18	2.3	2.85	2.41	2.70
2018Q2	29-Jun-18	4.1	2.79	2.73	2.77
2018Q3	28-Sep-18	3.5	2.47	3.87	3.32
2018Q4	28-Dec-18	2.6	2.48	2.70	2.69

2019Q1	29-Mar-19	3.2	1.31	2.22	2.07
2019Q2	28-Jun-19	2.1	1.30	2.31	2.25
2019Q3	27-Sep-19	1.9	2.06	3.19	2.94
2019Q4	27-Dec-19	2.1	1.19	1.98	1.94
2020Q1	27-Mar-20	-4.8	1.68	2.61	2.47
2020Q2	26-Jun-20	-32.9	-16.33	-4.36	-4.02
2020Q3	25-Sep-20	33.1	14.09	2.38	1.55
2020Q4	31-Dec-20	4	2.05	2.70	7.43
2021Q1	26-Mar-21	6.4	6.05	2.33	3.37