Machine Learning-Based Forecasting of Qatar's LNG Export Volumes Under Market Volatility

Madhava Miglani

Doha College, PO Box 7506, Al Niser Street, Doha, Qatar
Email: *madhava.miglani[at]gmail.com*

Abstract: ***The uses of Liquified Natural Gases (LNG) are vast, from generating electricity at huge scales for a town's grid to being burnt using a gas stove to cook. LNG is just Natural Gas (NG) cooled down to -162°C (-260°F) to turn it into liquid form to make it easier to transport as, the volume has decreased (meaning more can be transported at once) and that it is in a liquid form which is more stable making it less prone to major disasters. Qatar is a major exporter of LNG, in fact it is the third largest producer of it after the US and Australia. Qatar produces 20% of the global supply, this means many countries rely on Qatar for their LNG. The main goal of this study is to build statistical models which can accurately analyse and predict Qatar's LNG exports based on previous data and benchmark variables. The data used in this study consist of Qatar's monthly export values for the period 2019-2024, the export volume and values in terms of the mass exported from Qatar for the same period, and finally the Henry Hub and Asia LNG prices as gas price benchmarks. The statistical and machine learning models used in this study are the ARIMA and the Random Forest models. The results show that while ARIMA provides a useful baseline prediction, the Random Forest model performs better because it can include more variables such as global gas prices. These findings suggest that machine learning methods can improve forecasting accuracy for LNG export markets.***

Keywords: Qatar, LNG exports, machine learning, econometric models

1. Introduction

Natural gas (NG) is a combustible gas that is obtained from natural sources. It is made up of different chain sized hydrocarbons, however, 90% of NG is made up of methane (Sun-Feel Yang et al, 2023). This composition mainly varies based on the region it is extracted from and its refining processes. These refining processes are generally carried out to adjust the calorimetric properties. Generally, NG produced from gasfields undergoes a process to change its physicochemical properties and liquify it, to make it easier to transport and to make the end fuel more valuable. This process is carried out by cooling the NG (using cryogenic chilling) to below -161°C (Wood et al, 2017), this results in NG's volume decreasing by 1/600 compared to its gaseous state.

In the Fiscal Year of 2025 the world trading volume of Liquefied Natural Gas (LNG) was 411.24 million tonnes (MT) (IGU, 2025) of which 82 MT was exported by Australia making it the largest LNG Exporter in the world, it was closely filled by Qatar and USA as the world's second largest and third largest exporters of LNG.

LNG exports are affected by many factors, some of the main factors that affect global exports are geopolitical tensions, natural disasters, and widespread illnesses. For example COVID - 19 drastically affected the global LNG markets negatively, this widespread disease decreased the total trade volume's growth. This is evident as the growth between the previous years (between 2018 - 2019) was approximately 11%, whereas in the years following the pandemic (2020 - 2021) the growth rate dropped to around 2-3%, as industrial consumption and export volumes declined due to lowered demand and logistic restrictions (IEA, 2020). Furthermore, natural disasters such as earthquakes and hurricanes can disrupt natural gas supply chains and liquefaction facilities, leading to decreased export capacity and energy shortages in importing regions (BP, 2024). Another important factor that can affect LNG exports is geopolitical tension, which can impact trade relations, shipping routes, and market stability.

Over the last decade, the structure of the LNG market has evolved significantly. The emergence of new exporters such as the United States and the increased demand from Asian countries like China, Japan, and South Korea have diversified trade flows (IEA, 2024). This shift towards flexible and short-term contracts has made the LNG market more dynamic but also more unpredictable. These unpredictable changes highlight the importance of accurate forecasting models that can predict the export volumes of LNG.

Forecasting LNG export volumes is a complex process because it is influenced by factors such as production levels, regional gas prices, global demand, and weather conditions. Statistical and machine learning based models are commonly used to predict these fluctuations, each with its own set of advantages and disadvantages. A traditional model used for time series forecasting is the Autoregressive Integrated Moving Average (ARIMA) model, which uses historical data to identify patterns of trends and seasonality (Zhang et al, 2021). However, ARIMA assumes linear relationships and therefore may not perform well when there are nonlinear patterns or sudden changes caused by market disruptions.

In recent years, more complex models based on machine learning have been used to overcome these limitations. Algorithms such as Random Forest (RF) and XGBoost (Extreme Gradient Boosting) are able to capture complex nonlinear relationships between different variables such as production capacity, temperature variation, global GDP, and gas pricing indexes (Li et al, 2021). Moreover, deep learning models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) can model sequential data by learning long-term dependencies and temporal relationships in time-series data (Wang et al, 2024). Hybrid models that combine both statistical and machine learning methods, such as ARIMA-LSTM, have shown improved performance and accuracy as they integrate the strength of linear trend analysis with the flexibility of deep learning models (Chen et al, 2024). These models are particularly useful when the market experiences frequent disruptions or irregular patterns, such as those caused by geopolitical or climate-related factors.

Therefore, the use of models such as ARIMA, Random Forest, XGBoost, LSTM, and hybrid combinations offers a strong methodological basis for predicting LNG export volumes accurately. This can help exporting countries like Qatar, Australia, and the USA to optimize their production and manage supply schedules more effectively, enhancing their competitiveness in the global LNG market.

In this paper, we explore the factors that influence Qatar's monthly LNG export values and evaluate how well different forecasting methods can predict future export performance. To do this, we use monthly data from January 2019 to December 2024, combining official export statistics with international gas price benchmarks. The methods we used are ARIMA and Random Forest, both discussed in Section 2. The export data were obtained from Qatar's customs trade statistics and include total monthly export value (in QAR and USD), total export quantity, and total export weight for petroleum gases (HS4 code 2711). To capture global market conditions, we also include Henry Hub natural gas prices (U.S. benchmark) and Asia LNG prices, which reflect conditions in Qatar's main export markets. The final dataset consists of 72 monthly observations and includes both economic variables and lagged export values. Full descriptive statistics and visualisations are presented in Section 3. In terms of analysis (Section 4), we begin by assessing how well ARIMA predicts LNG exports using only historical export data, and we then extend the analysis by incorporating global price variables (Section 4.1). Following this, we apply a modern Machine Learning method, Random Forest, and evaluate its predictive accuracy in forecasting LNG exports (Section 4.2). Finally, we compare the forecasting performance of the traditional ARIMA approach and the Random Forest model in Section 4.3. A conclusion to the manuscript is available in Section 5.

2. Methods

2.1 ARIMA model

The Autoregressive Integrated Moving Average (ARIMA) model is a classical statistical approach used for time-series forecasting, where future values are predicted using patterns embedded in past observations. ARIMA is particularly suited to univariate time series (i.e., forecasting one variable such as LNG export volume over time) because it models how a variable evolves based on its own historical behaviour, rather than relying on many external predictors. In practice, ARIMA is widely used as a baseline forecasting method because it is mathematically interpretable and performs well when the underlying relationships are approximately linear and the data structure is stable over time.

ARIMA works by combining three components, represented as ARIMA(p,d,qp,d,qp,d,q). The autoregressive (AR) part (ppp) captures persistence in the series by modelling the current value as a function of previous values (lags). The integrated (I) part (ddd) refers to differencing, which is applied to remove non-stationarity (e.g., long-term upward/downward trends) so that the series has a more constant mean and variance across time. The moving average (MA) part (qqq) models short-run fluctuations by using previous forecast errors, allowing the model to correct for shocks that are not explained by the lagged values alone. After selecting ppp, ddd, and qqq, the model is fitted to the historical series and then used to generate forecasts by projecting the learned lag structure and error dynamics forward.

To perform ARIMA forecasting reliably, the main requirement is a sufficiently long and consistently measured time series (for example, monthly LNG exports measured in million tonnes across multiple years). The data should be collected at a fixed frequency (monthly, quarterly, etc.) with minimal missing values, because irregular spacing breaks the assumptions of standard ARIMA implementations. Since ARIMA assumes stationarity after differencing, the dataset must either already be stationary or be transformable into a stationary series using differencing (and sometimes variance-stabilising transformations such as logarithms when volatility increases with the level of the series). In addition, ARIMA is most effective when the series reflects a relatively stable data-generating process; if there are major structural breaks (e.g., sudden policy shocks, pandemics, or geopolitical disruptions), the model may require re-estimation, intervention terms, or extensions (such as seasonal ARIMA) to maintain forecasting accuracy.

ARIMA models are predominantly used in fields where data is naturally ordered over time and decision-makers care about short- to medium-term forecasts. They are widely applied in economics and finance, for example to model and forecast variables such as GDP, inflation, interest rates, exchange rates, and stock indices, because these series often exhibit trends that ARIMA is designed to capture. In the energy sector, ARIMA is frequently used to forecast energy demand, electricity load, natural gas consumption, and commodity export volumes, since these time series are typically reported at regular intervals (daily, monthly, or yearly) and are influenced by historical patterns and seasonality. Beyond economics and energy, ARIMA is used in transportation, environmental science, and public health to forecast variables such as traffic flow, pollution levels, and disease incidence for planning and resource allocation. It is also applied in business analytics for sales forecasting, inventory control, and call-centre workload prediction, where its simple and interpretable structure allows modelling of linear temporal dynamics without many explanatory variables.

ARIMA forecasting is implemented in several widely used R packages that provide both core functionality and convenient helper tools. The base implementation is available in the built-in stats package via the arima() and predict() functions, which allow you to fit custom ARIMA models by specifying the p, d, and q orders directly. For more automated and user-friendly workflows, the forecast package (by Hyndman and colleagues) offers auto.arima(), which selects appropriate model orders based on information criteria and supports seasonal ARIMA and stepwise search.

Paper ID: SR26307185914 DOI: https://dx.doi.org/10.21275/SR26307185914

2.2 ML Random Forest

Machine learning is a type of artificial intelligence that focuses on building models which learn patterns from data instead of being explicitly programmed with fixed rules. In practice, a machine learning algorithm is given historical data called training data (inputs together with the correct outputs) and adjusts its internal parameters so that it can make accurate predictions or classifications on new, unseen data, i.e., predict unseen behaviour. This data-driven approach is widely used in tasks such as forecasting, image recognition, and fraud detection, where writing hand-crafted rules would be impractical or impossible.

Random forest is a supervised machine learning algorithm that builds a large number of decision trees and then combines their outputs to produce a single, more robust prediction. Each tree is trained on a slightly different random subset of the training data and, at each split in the tree, only a random subset of features is considered; this randomness decorrelates the trees and helps reduce overfitting. For classification problems, the random forest predicts the class that most trees vote for (majority voting), whereas for regression problems it typically predicts the average of the individual tree outputs. Because of this ensemble structure, random forests tend to handle noisy data well and often achieve strong performance with relatively little parameter tuning.

To evaluate how well the random forest generalises, the available dataset is divided into a training set and a testing set. The training set is used to fit the model: the forest learns patterns and relationships from these observations. The testing set is kept completely separate during training and is only used at the end to assess performance on unseen data, providing an unbiased estimate of how the model is likely to behave in real-world use. In many applications, a common choice is to allocate around 70-80% of the data to training and the remaining 20-30% to testing, but the exact split can be adjusted depending on dataset size and research design.

3. Data

The export data used for the analyses in this study were obtained from Qatar's customs trade statistics. The data include the total monthly export value (in QAR and USD), total export quantity, and total export weight for petroleum gases (HS4 code 2711). To capture the global market conditions, we also account for international gas price benchmarks, specifically the Henry Hub natural gas prices (U.S. benchmark) and the Asia LNG prices. Both benchmarks are likely reflecting conditions in Qatar’s main export markets. The dataset consists of 72 monthly observations, i.e., period spanning 6 years 2019-2024, and includes both economic variables and lagged export values. This data can be visualised in Figure 1.

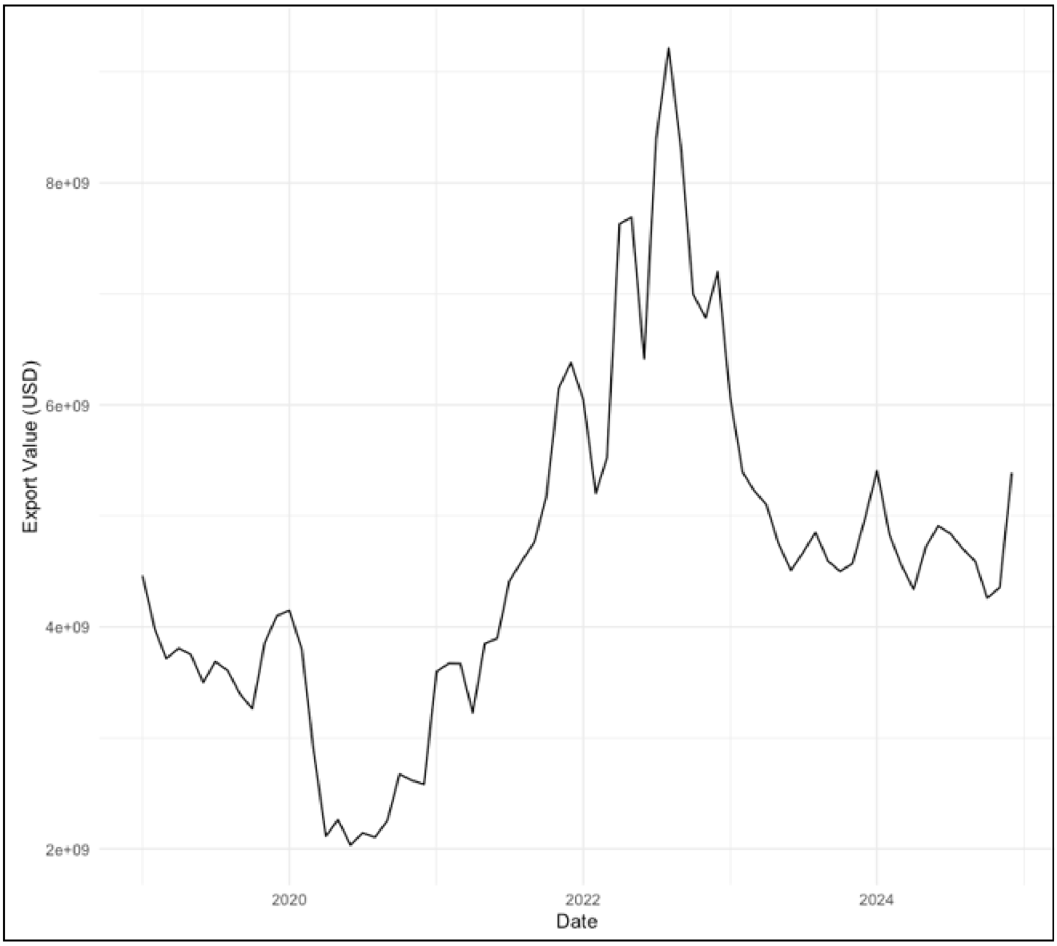

Figure 1: Time series plot of the export volume of Qatar’s LNG Exports for the period 2019-2024.

It is important to acknowledge that these data reflect recent years where there have been major events such as global pandemics, wars and tariff battles. Therefore, there is a great amount of fluctuation. This is clearly noticeable in the summary of data. For instance, the maximum export value was 9.2 billion USD (16.7 billion QAR) and the minimum was 2.0 billion USD (7.4 billion QAR), which shows that there is a difference of 7.2 billion US dollars or 9.3 billion QAR. From this range we can learn that there are many factors that can gravely affect the export volumes of LNG from Qatar. This can clearly be noticed specifically in the year of 2022 (last two quarters mainly) where due to the global energy crisis long term customers of other markets (e.g., Australia, USA, etc.) shifted and bought cheaper energy from Qatar, majorly increasing net export volumes for that year. Furthermore the mean export value for Qatar was 4.6 billion and the median was 4.5 billion dollars, showing how reliant Qatar's 179 billion (according to 2021 stat) USD economy is on LNG.

Our data also provide export volumes in terms of weight and volume of gas. Summarising the data we also see a spike in the mass (kg) exported in the year of 2022, however, there is a greater difference between the mean of price and the highest spike in 2022 (it was 4.7 billion USD) whereas the increase for the mass was a whopping (6.4 billion Kg). This suggests large value changes are driven more by price than volume. This means that price factors affect the volatility more than non-price factors.

The data that will be used to train and develop the models in the next section will have 13 variables in total. The base will be the LNG export value data, however, there will be few additional variables to improve the predictive accuracy. These variables are: the date, the export value in Qatari Riyal (QAR), the total quantity (as in volume), total weight in kilograms, the export value in United States Dollar (USD), the year of the month, the month of the year, the quarter of the fiscal year, Lag1 value (value of the previous month), Lag3 value (values of export for the previous 3 months), MA3 Value (the moving average of the past 3 months), the prices of LNG in the HenryHub market and in the Asian market.

4. Results

In this section, we analyse Qatar’s monthly LNG export values using different forecasting methods. We first apply a traditional ARIMA model to examine how well past export values can predict future exports. We then compare these results with more advanced approaches like Random Forest to evaluate whether including additional information improves forecasting accuracy.

4.1. ARIMA analysis

The ARIMA Models used to predict the export volumes of liquified natural gas from Qatar are either univariate or multivariate models. In the univariate model we use the only variable the model is given is the past exports from Qatar, this causes the model to be more inaccurate as there are no more variables to help predict the exports of the next years and months. This causes the mean absolute percentage error (MAPE) to be higher as there aren't many variables the model can base its prediction off of, this means it can only analyse general trends, e.g that there the export volume is greater near the end of the quarter compared to the start (may not be true just for example). As the model only has one variable it can analyse only basic patterns it cannot take any other factors such as aggregate supply in the market for LNG this causes the predictions to be inaccurate by a lot in this case the MAE for the univariate model is 7.90%. A MAPE of 7.90% means that on average, your forecast error is about 8% off actual values, this value of MAPE is reasonable for as the energy markets are extremely volatile and furthermore due to the 2022 anomalous shock period the data is not that consistent making it harder for the model to create patterns leading to more error, based on this we can conclude it is a good baseline model.

Now to decrease the MAPE of the model we would add more variables, in our study we add the export data for the Henry Hub and Asian LNG export values, we do this so that the model can access the aggregate supply for the LNG markets. In theory this should greatly benefit the model’s accuracy, this is because if the model can see the aggregate supply, it can decide if the export volume is going up or down. For example if the aggregate supply (from countries/regions apart from Qatar) goes down the export volume from Qatar will most likely go up as those who would buy from other producers may shift and buy from Qatar as the volume produced by others is not enough. This works not only in theory but in reality as well, the main example of this is 2022, this is because there is a huge spike in the volume exported as there was an energy crisis in 2022 where the exports of The Henry hub fell, this meant those importing from alternate markets shifted to Qatar to import. From this we can conclude that as the number of columns increases the MAPE goes lower increasing the accuracy of the models predictions.

4.2 Random Forest analysis

Random Forest, a machine learning approach, was applied to create a model which can accurately predict Qatar’s monthly LNG export values. Unlike ARIMA, which mainly relies on past export values and statistical structure, Random Forest is a machine learning method that can capture complex and non-linear relationships between variables, this means that the model can identify more complex patterns which can increase predictive accuracy, and hence, cause MAPE to decrease. The Random Forest model was trained on data from 2019-2023 (referred to as the training dataset) and tested on 2024 data (referred to as the testing dataset) to evaluate its predictive performance. The actual and predicted LNG export values for the period 2024 are shown in Figure 2.

Paper ID: SR26307185914 DOI: https://dx.doi.org/10.21275/SR26307185914

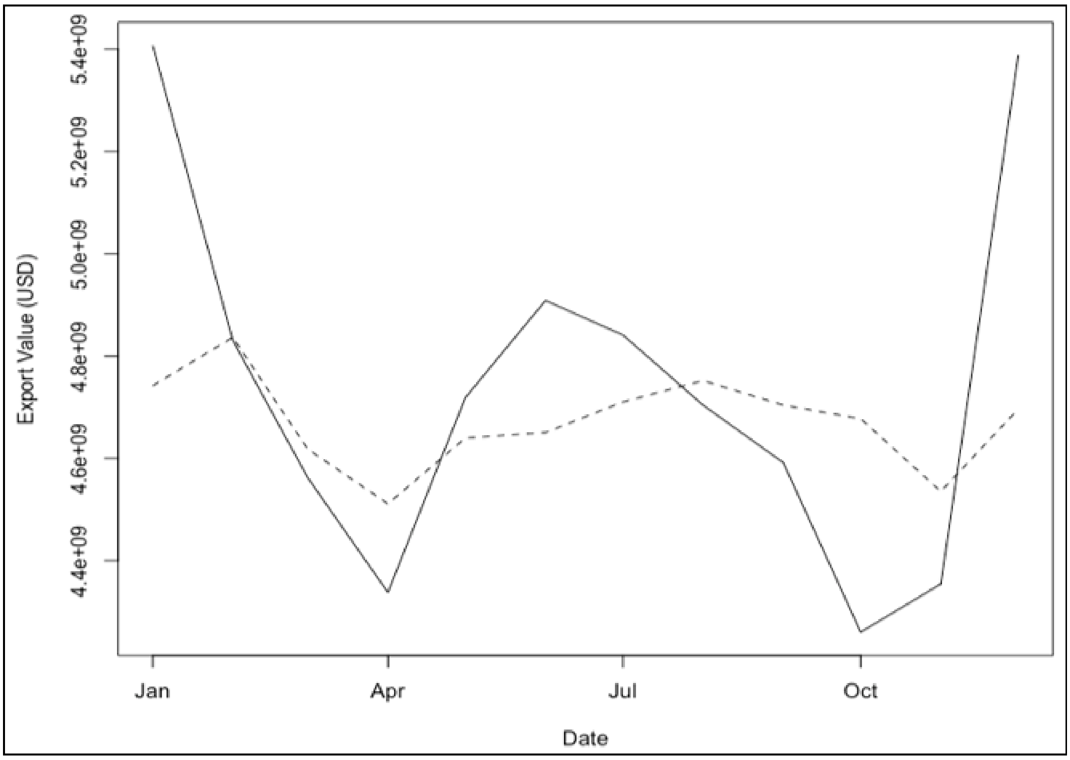

Figure 2: Actual vs predicted LNG export values for the 2024 test period using the Random Forest model. The solid line represents the actual monthly export values, while the dashed line shows the values predicted by the model. The plot shows that the Random Forest model captures the overall trend of exports reasonably well, although it slightly smooths some of the larger fluctuations.

The Random Forest model explains approximately 93.6% of the variation in the training data, which indicates a very strong fit. When evaluated on the test data, the model achieved a MAPE of 4.79%. This means that, on average, the model's predictions were only about 4.79% away from the true export values. Compared to the univariate ARIMA model (≈7.9% error) and the multivariate ARIMA model (≈5.1% error), the Random Forest shows the lowest prediction error, hence, better predictive accuracy. However a question arising is why the RF model is praised for its accuracy even though it cannot accurately forecast the extremes. This is because the model finds patterns that are constant therefore it is unable to predict extremes fully. Furthermore, this model is considered as a good fit model as it can predict the general pattern, i.e., downward or upward trend of the LNG export values. This is an expected and reasonable behaviour for applications like the energy markets which are considered in this study. The energy markets are very volatile due to the number of variables affecting them, i.e., cost variables, and also the large number of exogenous factors affecting it, i.e., economic or even natural disasters. Due to the latter statement, it is very challenging to build a more accurate model because access is needed for a larger number of factors. Hence, given the current availability of data the random forest model is considered to have a high predictive accuracy.

In contrast to the ARIMA model, the Random Forest model considered the maximum amount of variables we had in hand. These new information variables may have been a contributor to the increased accuracy of the random forest. Specifically, in the RF model we added a Moving Average value across 3 months (MA3_Value) which corresponds to the average of the past three months; one month lag value (Lag1_Value) which contains the value of the previous month's exports; Henry Hub, this variable is used as a comparative benchmark as the Henry Hub is the market of LNG from The United States of America; Asian_LNG_Price, its use case is the same as the Henry Hub Variable, as this contains the value of the exports by the asian market; Lag3_Value, this column of data contains the export values for the past three months; Total_weight, is the variable which has the details of the actual mass of LNG that is exported out of Qatar; Month just displays which month the exports are for; total_quantity, gives the model the volume of the export; Quarter, just displays the quarter of the fiscal year this month was in. Due to the huge increase in the number of variables and the quality of variables the Random Forest (RF) model out performs the ARIMA by a great margin.

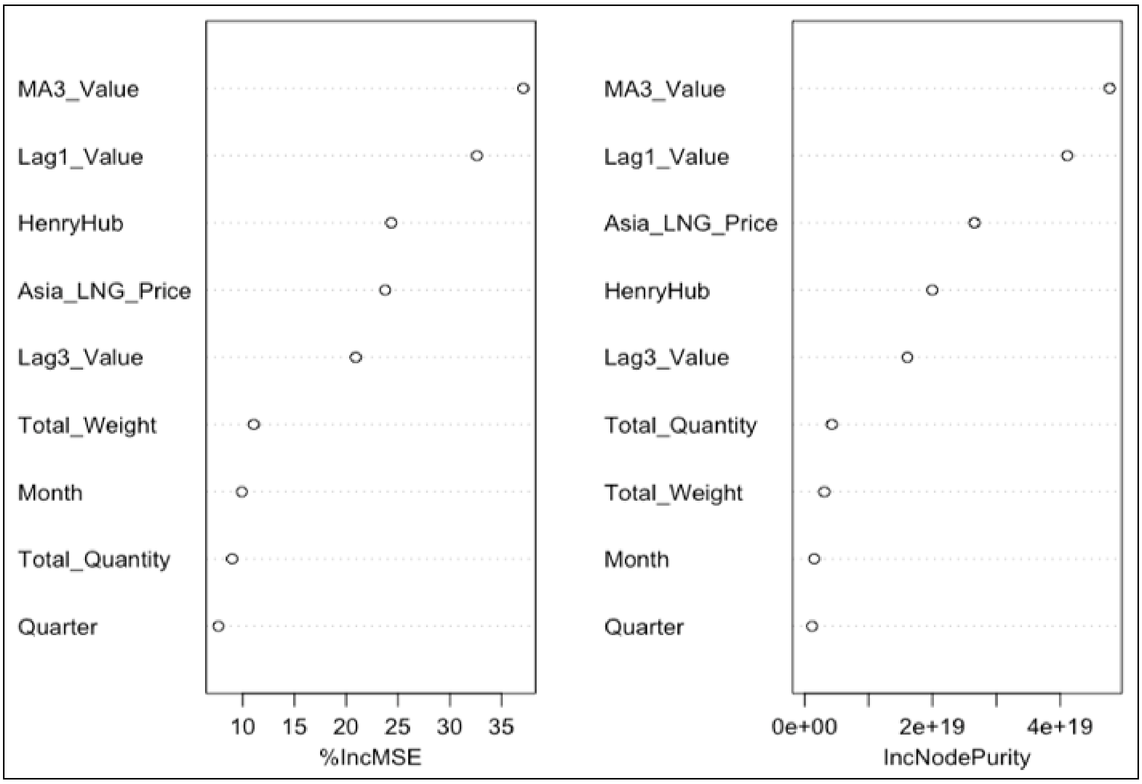

Figure 3: Variable importance plot from the Random Forest model representing the most impactful variables used to predict LNG export values. The left panel (%IncMSE) shows how much the model's prediction error increases when each variable is removed, indicating its importance for accurate predictions. The right panel (IncNodePurity) shows how much each variable contributes to improving the model's decision trees. The results suggest that the recent export history (Lag1_Value and MA3_Value) and global gas price benchmarks (Henry Hub and Asia LNG Price) are the most important factors in predicting LNG export values.

The variable importance plot shows that MA3_Value (three-month moving average) and Lag1_Value (last month's exports) are the most important predictors. This means recent export history plays a major role in predicting future exports. Among the external economic variables, Henry Hub price and Asia LNG price are also important drivers. This confirms that global gas price movements strongly influence Qatar's LNG export value. In contrast, variables such as month, quarter, and total quantity were less important in the model. The importance plot was based on %IncMSE and IncNodePurity. IncMSE% (the plot on the left) is the more important measure we should look at, IncMSE% means how much worse does the model perform if we randomly shuffle this variable? If there is a variable that is shuffled and there is not much change in the accuracy of the predictions that means that the variable is not very important or it does not have much impact on the predictions meaning that the variable can be replaced with any other variable to increase the overall accuracy of the model. The plot on the right shows the importance based on IncNodePurity, this is a variable importance measure that is used in RF models that quantifies the amount a predictor improves the purity of decision trees split across the forest of decision trees.

4.3 Model comparison

Based on our findings by creating two variants of an ARIMA model (univariate and multivariate) and creating a RF model and feeding these models with data about liquefied natural gas exports from Qatar that ranges between 2019-2025, we learn that a multivariate model works much better as compared to a univariate model. This is clearly seen as the MAPE for an Univariate model was 7.90% whereas the MAPE for a Multivariate model was only 5.13%. Furthermore, a traditional machine learning model such as the random forest performed better than the ARIMA models with a MAPE of only 4.79%. This shows clearly that a classic machine learning model is better than a statistical model (an ARIMA model), this is mainly because the RF model can identify patterns more easily as compared to a statistical based model.

5. Conclusion

The goal of this study was to build models that can accurately predict the LNG export values of Qatar. To achieve this goal we used two types of models, an ARIMA model and a machine learning Random Forest model. In this study we used a univariate and a multivariate ARIMA model. A univariate model is only given one variable and uses that to identify patterns and predict, unsurprisingly this is not the best as it cannot see any relationships between variables as it only has one. Contrastingly, a multivariate model is given many variables to help it predict, in this case a multivariate model is much better as it can identify more complex economic relationships between the variables. However a Random Forest model outperformed ARIMA, providing evidence that machine learning algorithms can predict more accurately opposed to an ARIMA model. Furthermore this is because a RF model can spot non-linear trends and relationships between variables, making it more accurate at predicting the export volumes of Qatar. Based on our findings from this study we can conclude that a Random Forest model will be better for predictions for LNG exports as it outperforms both types of ARIMA models.

Predicting LNG markets will always be difficult as they are extremely volatile, due to the fact that these markets are extremely fragile and are affected extremely easily. This can be observed presently (as of March 2026) when the Islamic Republic of Iran closed the straight of Hormuz caused the prices of LNG to shoot up as 20% of global supply was killed (as Qatar can only export through there), this goes to show that there are many factors that influence the exports and price of LNG making it near impossible to get a a model that can perfectly predict the exports from Qatar. However the accuracy of these models can be increased by getting a larger dataset or by adding more economic and geopolitical variables to allow better insight into what causes the volatility in exports and prices. This is a future research objective.

References

[1] Abdul-Salam, Y., & Abdul-Salam, F. (2025). Qatar's LNG exports: Advancing efficiency in electricity generation and reducing carbon emissions in the global energy transition. *Energy Strategy Reviews*, *62*, 101978.

[2] Sun-Feel Yang et al. (2023). Composition of natural gas hydrocarbons. *Journal of Energy Resources*, 45(2), 123-135.

[3] Wood, D. A., et al. (2017). LNG liquefaction processes and volume reduction. *Journal of Natural Gas Science and Engineering*, 40, 1-15.

[4] IGU. (2025). World LNG Report 2025. *International Gas Union Annual Report*, 1, 1-50.

[5] IEA. (2020). Global energy review 2020: Impact of COVID-19. *International Energy Agency Reports*, 2020, 45-60.

[6] BP. (2024). Statistical review of world energy. *BP Energy Outlook*, 2024 Edition, 100-120.

[7] IEA. (2024). LNG market review 2024. *International Energy Agency LNG Analysis*, 2024, 20-35.

[8] Zhang, G., et al. (2021). ARIMA models for LNG forecasting. *Energy*, 225, 120287.

[9] Li, X., et al. (2021). Machine learning for LNG demand prediction. *Energy Reports*, 7, 102-115.

[10] Wang, J., et al. (2024). LSTM and GRU in energy time series. *IET Renewable Power Generation*, 18(5), 12948.

[11] Chen, Y., et al. (2024). Hybrid ARIMA-LSTM for LNG exports. *Energy*, 292, 130282.