International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(10s) 2026 pp. 1204-1223
https://doi.org/10.70917/ijcisim-2026-3709
Article

Explainable Machine Learning for Wood Market Forecasting Using News Analytics and Temporal Causality

Poonam Gohil[1], Sheshang Degadwala[2]

[1]Department of Information Technology Engineering, Indus University, Ahmedabad, Gujarat, India
Email: poonamgohil3892@gmail.com
[2]Department of Computer Engineering, Sigma University, Vadodara, Gujarat, India.
Email: sheshang13@gmail.com

Abstract: The ability to forecast accurate dynamics of the wood market is vital for the management of the supply chain, production planning and sustainable forestry decision making. Most of the conventional techniques for forecasting are based on numerical data from the past and do not take into account the predictive capabilities of unstructured text data available from online news sources. In this research, an explainable framework based on news data and machine learning is proposed to predict wood sales on weekly basis, which incorporates natural language processing, Uniform Manifold Approximation and Projection (UMAP), Granger causality analysis, LightGBM and SHapley Additive exPlanations (SHAP) algorithms. News articles related to wood were gathered and preprocessed to create the semantic embeddings, followed by UMAP dimensionality reduction and clustering into semantic clusters. Using granger causality analysis, statistically significant news derived features were identified and then used in conjunction with historical sales variables to train forecasting models: Random Forest and LightGBM. Based on the experimental results, the proposed LightGBM model had an MSE of 6.54, RMSE of 2.56, MAE of 1.50, and an R^2 value of 0.99, which demonstrated the model showed the excellent predictive accuracy and generalization ability. Moreover, the SHAP-based explainability showed the most influential historical and semantic features that influenced forecasting performance, offering more transparency and interpretation of the decisions to be made. Overall, the proposed framework shows the potential of combining semantic news analytics with explainable machine learning to improve the accuracy and interpretability of wood market forecasting, thereby providing a valuable tool for researchers, policymakers, and stakeholders in the forestry and wood products sector.
Keywords: Wood market forecasting; News analytics; Explainable artificial intelligence (XAI); LightGBM; Time series prediction.

1. Introduction

The wood industry is an important contributor towards sustainable economic development, sustainable use of renewable resources, and the concept of circular bioeconomy. Precise wood production, demand and market forecasting is therefore vital for optimizing wood supply chains, better wood inventory management, minimizing the cost of wood operations, and informed strategic decision making. In recent years, a growing number of research has directed towards predicting the availability of wood and its circularity through quantitative modeling approaches, as an aid for sustainable resource management approaches [1]. At the same time, recent progress in machine learning has boosted the accuracy of forecasting in various supply chain applications and shown to outperform traditional statistical methods for dealing with nonlinear relationships and complex temporal dependencies [2]. Ensemble learning methods, such as random forest and gradient boosting have also been used to predict yields of forest products under different climatic and socio-economic scenarios [3] and panel econometric and machine learning have been successfully used to analyse the determinants of wood pulp production and industrial output [4]. In addition, there are some models that predict demand for wood products in resource-limited countries, which shows the significance of having reliable demand predicting models for planning the market and formulating policies [5]. Forecasting methods are still in a rapid development stage in different industrial sectors [6] and there are still considerable problems in the wood industry caused by market fluctuations, uncertainties about the environment and external socio-economic events.

Artificial intelligence and deep learning are increasingly used to predict in the forestry, logistics and wood product sectors, as shown in the latest literature. A comparative study evaluated the applicability of classical statistical, machine learning and hybrid forecasting models for wood industry applications and showed a general trend towards using data-driven models [7]. The use of machine learning models often surpasses traditional time-series models in cases where nonlinear interactions prevail in the market, as demonstrated by simulation studies [8]. ANNs have been applied to model the bilateral trade flows of wood [9] and stochastic optimization has been applied to enhance the wood pellet supply chains, incorporating demand forecasting [10]. In addition to this, several other machine learning advances in forecasting have been shared across other industries, such as manufacturing, automotive, textile, furniture and ecological systems, which demonstrates the widespread applicability of machine learning for demand prediction and decision support [11], [12], [18], [20], [22]. In recent years, Long Short-Term Memory (LSTM) networks, graph neural networks, and multi-source data fusion approaches have shown outstanding performance in making forest product trade and timber price forecasts, leveraging the temporal relationships among the underlying data sources [13] and [15]. The need for smart forecasting systems for wood product chains and international trade that can respond to the fast changing market conditions is further highlighted in comprehensive reviews of global wood product supply chains and international trade [16; 17].

Notwithstanding these developments, current forecasting research focuses mainly on past production, economic, climatic or trade data, and generally ignores the predictive signals in unstructured, text-based data like online news. There is a good chance that news coverage will reflect new developments in the world economy, new policies, new supply disruptions, new environmental events, and new industrial developments before they are evident in traditional numerical indicators. Thus, the use of natural language processing (NLP) in conjunction with time-series prediction holds the promise of increased prediction accuracy as well as model interpretability. In this study, we suggest an interpretable news-based machine learning model for predicting the dynamics of the wood market using news analytics, semantic feature extraction, temporal causality analysis, and ensemble learning models. The proposed methodology is based on the following steps: keyword analysis, latent semantic clustering using the Uniform Manifold Approximation and Projection (UMAP), Granger causality analysis and the supervised learning model using either Random Forest or LightGBM, followed by explainability using SHAP. The proposed framework combines structured market information with unstructured news information, creating both accurate forecasts and interpretable insights into the drivers of wood market behavior that can guide evidence-based decisions on sustainable management of forest products, industrial planning and resilient wood supply chains.

2. Related Works

Given the growing need for the world to consume forest products, forecasting has become a key element in sustainable wood resource management and optimizing the forest supply chain. More recently, a predictive modelling of wood availability, production and consumption has been the subject of studies, aiming to support the implementation of circular economy activities. A major research topic for sustainable resource use is the development of predictive models like the PRecTimber 2.0 framework presented by Mahenthren et al. [1] to predict quantitative and qualitative wood circularity potentials on a national level (Germany, Finland and Sweden). Bhavikatta and Goluguri [2] analyzed Learning algorithms such as Long Short-Term Memory (LSTM) networks for forecasting in Sustainable Supply Chain with multiple datasets, which showed that the deep learning network with the ability to capture temporal dependency has resulted in better forecasting accuracy. Likewise, ensemble learning, such as the Random Forest algorithms used by Chen et al., [3] proved to be an effective approach for modelling nonlinear environmental interactions by predicting forest product yields in China based on combination of climatic and socio-economic variables. Aydogan [4] applied the panel econometric analysis and machine learning to bring the factors behind the chemical wood pulp production in Europe to light, showing that hybrid explanatory methods can be quite effective for explaining the dynamics of industrial production. In Tunisia, with little forest resources, Daly-Hassen et al. [5] developed demand and net import forecasting models for wood products and showed that well-designed forecasting models can have significant impacts in the policy-making and strategic planning processes. All of these studies confirm that machine learning increasingly can be used to predict wood production and consumption but are mostly based on numerical data that can be structured and are lacking in the use of unstructured textual data that might offer early signals of market fluctuations.

Many researchers have made comparisons between forecasting algorithms in order to identify the most suitable models for wood industry applications and related supply chains. Turkay et al. [6] proposed scenario-based forecasting models for predicting future energy consumption and carbon footprints of AI, highlighting the need for accurate long-term forecasting for sustainable development. Biernacka [7] discussed different forecasting models applied in the wood industry, and he found that none of the models has proven to be universally superior in different operational conditions, indicating that the quality of the forecast depends to a large extent on the quality of the data used and the application fields. Schmid et al. [8] did a thorough simulation study to compare various statistical forecasting methods and machine learning algorithms for the prediction of logistics time-series and found that in the presence of nonlinear patterns, machine learning models generally outperform traditional statistical forecasting methods. Morland et al. [9] examined two models, artificial neural networks and gravity models, for the forecasting of bilateral wood trade and found that the artificial neural networks model forecasts with greater capability for complex international wood trade. In the field of wood pellets, Vazifeh and Mafakheri [10] demonstrated the value of predictive modelling combined with operational optimization by employing stochastic optimization to enhance wood pellets supply chains. Moreover, global trends on the forecasting methods in the automobile industry [11] and the textile and fashion industry [12] support the relevance of AI in industrial forecasting. In general, all these investigations have shown that data-driven forecasting is now essential in various industrial applications, but in some limited way, the integration of heterogeneous information sources, like textual news data and traditional numerical variables, has also been little explored.

The application of artificial intelligence (AI) has greatly broadened the scope of deep learning and hybrid machine learning approaches in the fields of forestry and timber markets in recent years. To predict cross-border forest product trade, Yang and Zhang [13] proposed an LSTM based forecasting framework, which also adopted multi-source data fusion to achieve better forecasting accuracy, revealing that the fusion of multiple information sources is a significant improvement in forecasting accuracy. Huang et al. [14] reviewed previous studies and future research directions on timber structures, and used time-series forecasting methods to forecast the future growth of the timber industry. Thao et al. [15] proposed a hybrid GNN and LSTM network to predict timber prices in Vietnam, which achieved excellent performance over traditional machine learning algorithms. Mensah et al. [16] provided an exhaustive review of the global wood products supply chain, and highlighted the growing complexity of international timber markets and the need for smart forecasting tools. Likewise, Upadhyay et al. [17] examined the production, import and export dynamics of the G20 countries, which gave useful insights into the dynamics of wood products trade in the global arena. Güngör et al. [18] also showed that multivariate time-series models can be used effectively for furniture demand forecasting, as they can include more than one economic indicator at a time. In addition to the forecasting application, the machine learning method has been successfully applied to sustainable forest management [19], ecological forecasting [20], predicting wood product properties during manufacturing [21], manufacturing demand forecasting [22], international trade forecasting [23], sustainable development assessment [24], prediction of retail sales [25] and long-term wood consumption forecasting [26]. These studies validate the broad applicability of machine learning in forestry and industrial applications, while also emphasizing the need for readily explainable predictive models that can aid decision-making processes in practical applications.

Previous forecasting research has paved the way for the use of machine learning in the forestry and wood product sector. Mathematical modelling techniques were applied by Sushko and Plastinin [27] to forecasting the market prices of forest products and the value of the price predictions for forest product industrial planning. Nepal et al. (2018) explored the impact of growing demand for mass timber on forests and wood product markets across the globe and the need for reliable forecasting tools that can help enable sustainable forest management. Tigas et al. [29] was one of the early applications of the use of an artificial neural network in the field of modelling wood consumption, showing that the artificial neural network models are superior to some traditional models in modelling the nonlinear consumption behaviour. Likewise, Anandhi et al. [30] used artificial neural networks to predict pulpwood demand and supply, showing that AI models can be used to model complex market dynamics, even with relatively small datasets. All these seminal studies proved the utility of the intelligent forecasting models, but they were mostly applied to structured numerical variables like production data, consumption data, pricing data and trade volumes. With the progress of natural language processing, semantic feature extraction, explainable artificial intelligence, and transformer-based text analysis, introducing text-based data has gained new prospects into early trend identification and enhanced the understanding of prediction outcomes.

While great strides have been made in the forecasting of forestry and wood product markets using machine learning, there are still a number of research gaps in the current literature. First, most of the previous studies are highly dependent on structured numerical variables, such as production figures, climate indicators, trade volumes and economic variables, with very little attention given to the massive amount of information to be found in news articles and on the Internet that is unstructured. Second, while deep learning and ensemble learning methods have shown to have high predictive performance, relatively few studies consider whether there are statistically significant causal relationships between news events and future wood market behaviour when adding them to forecasting models. Third, most of the currently available forecasting frameworks focus on the level of prediction accuracy, while they offer limited model interpretation that means they are not very useful for the industrial stakeholders and policy makers who need to understand the prediction results. Fourth, only a few studies integrate semantic text mining, dimensionality reductions, causal inference, machine learning, and explainable artificial intelligence in an integrated forecasting model. Lastly, an integration of news analytics, semantic clustering, temporal causality analysis, ensemble learning and SHAP-based explainability has not been studied yet for forecasting the wood markets. To overcome these issues, this study introduces an explainable news-driven forecasting framework that integrates NLP, UMAP-based semantic clustering, Granger causality analysis, Random Forest, LightGBM and SHAP interpretation, which aims to enhance the accuracy of forecasting and offer clear insights into the drivers behind the wood market dynamics. The proposed framework brings in a comprehensive, interpretable and data driven methodology that can contribute to sustainable forest management, smart supply chain planning and evidence-based decision making in today's wood markets.

3. Methodology

This research presents an all-encompassing approach for using online news analytics and machine learning to predict the dynamics of the wood market with explainable artificial intelligence (XAI). In contrast to traditional forecasting methods, which use only numerical historical data, the proposed approach utilizes text data derived from wood-related news articles to adapt to new market trends, policy announcements, new developments in the industries, and wood supply chain disruptions. Overall, the workflow involves a series of steps, such as data collection, data preprocessing, NLP, semantic feature extraction, temporal causality analysis, machine learning-based forecasting, and model interpretation. The over-all research framework used in this study is shown in figure X. The goal is to provide a more accurate forecast and explainable artificial intelligence to reveal the factors that impact wood market behavior that come from news.

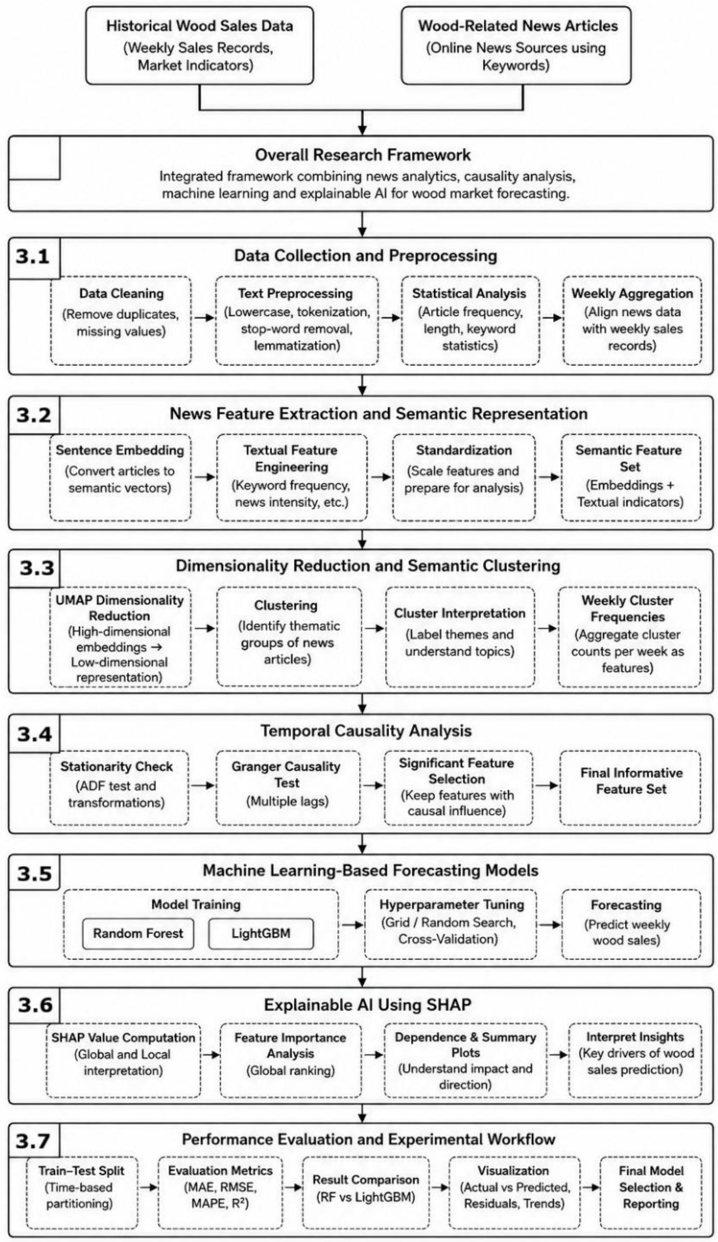

Figure 1. Proposed Methodological Framework of Wood Market Forecasting

The methodology starts by gathering historical wood sales data in addition to news articles from online sources that cover wood, forestry, wood products, construction materials, trade and sustainability. To obtain clean textual data for further processing, the collected news corpus is subjected to several pre-processing steps such as removal of duplicate documents, handling missing values, normalization of text, removal of punctuation markers, tokenization, stop-word removal and lemmatization. A statistical summary is then created to explore article distribution, publication frequency, keyword counts and temporal coverage. Sentence representations from processed news corpus are then extracted using sentence embedding techniques, followed by dimensionality reduction with Uniform Manifold Approximation and Projection (UMAP) to retain the semantic similarities between the sentences while reducing the dimensionality of the features. The semantic vectors are then reduced and grouped together to form dominant thematic structures for different categories of wood-related news.

Semantic feature extraction is conducted to gain insight into the temporal relationships between news-derived variables and historical wood sales data, by means of Granger causal analysis to check if there are any statistical significant predictive power with respect to the specific news topic. These identified informative features are then merged with historical sales features and rolling statistical features to create the final forecasting set of data. We have trained two supervised machine learning algorithms: Random Forest and LightGBM to predict future wood sales per week. Model performance is assessed based on typical regression statistics and comparing actual and model values graphically. Lastly, SHapley Additive exPlanations (SHAP) are used to measure the importance of a feature and to give a clear explanation of individual predictions. This holistic procedure allows for precise forecasting, and it also offers meaningful analyses of the impact of semantic news data on the dynamics of the wood market, which can help industrial actors and market stakeholders in making informed decisions.

3.1 Data Collection and Preprocessing

The first part of the proposed framework is to gather structured numerical information and unstructured textual information and create a comprehensive forecasting set of data. The weekly data from historical wood sales records were used to represent market demand for the study period. Concurrently, news articles on wood were collected from publicly available online news sources with a set of pre-established keywords with a relevance to timber, forestry, lumber, construction materials, wood products, trade policies, supply chains, sustainability and environmental regulations. The metadata that had been added to each news article enabled the textual data to be matched with the weekly sales observations, which were also recorded as integer data.The metadata, including publication date, title, article content and source information, enabled the textual data in each news article to be synchronized with the corresponding sales observation per week, which had also been recorded as integer data. The news articles were generated continuously, and were from multiple publishers, therefore, there were initially duplicate records, inconsistent formatting, incomplete news article entries and irrelevant data in the collected dataset. To ensure quality and consistency of the data, a systematic preprocessing procedure was introduced which was followed before extracting features and building models.

The textual preprocessing pipeline was a sequence of operations that were applied to the raw news articles to transform them into a machine-readable format. Articles and records with missing critical information were duplicated and deleted to prevent data redundancies and better data reliability. Next, all the text was translated to lowercase to ensure lexical uniformity, and punctuation, special characters, numerical characters that do not have many semantic features, and web hyperlinks were removed. It was then tokenized to segment the articles one by one into lexical units, and the stop words were removed to remove the words which do not contribute much semantic information. Lemmatization was then used to convert words to their canonical root forms to ensure the uniformity of words in the vocabulary, while maintaining their original meaning in the context. These cleaned documents were then further processed to calculate some descriptive statistics such as article frequency, document length distribution, trends in publications, and occurrence frequency of keywords. News articles were aggregated on weekly basis to match the timeliness of the wood sales data set. The data noise reduction and preservation of meaningful semantic information through this integrated preprocessing approach lay the groundwork for reliable subsequent NLP, semantic feature extraction, and machine learning-based forecasting.

3.2 News Feature Extraction and Semantic Representation

After cleaning the news corpus, the semantic information was extracted from the preprocessed data, and unstructured textual data was converted into quantitative features for machine learning. The conventional methods of representation for text, such as bag of words and term frequency approaches, do not reflect word-to-word and word-to-document relationships, especially when these relationships are context-specific. Thus, in the proposed framework, the sentence-level semantic embeddings are used to produce dense vector representations which retain the meaning of the full news article. These processed articles were each mapped to a high dimensional embedding vector in order to be able to represent articles that were semantically similar to each other in feature space, despite varying writing styles and word use. This semantic representation helps with a more detailed description of news content, and allows the forecasting model to learn more complex relationships between market events and wood sales.

Semantic embeddings were then generated and other descriptive textual features computed to describe the news corpus. These comprised keyword frequency distributions, article length statistics, document frequency and number of publications per week. The Word frequency analysis was carried out to extract the most frequent terms across all the data, and the Temporal aggregation was done to determine the intensity of news for each week in the wood sales data set. The semantic vectors and statistics were then standardized to remove scale differences between vectors and statistics. The features as a whole characterize both the meaning and the publication aspects of the gathered news articles. The proposed framework combines the contextual embeddings with statistical signals from the text to capture latent market signals that will not be available from the numerical sales records. This means that the resulting feature set is a complete description of the market information and downstream machine learning models can use it to find complex relations between news events and future wood market behavior.

3.3 Dimensionality Reduction and Semantic Clustering

Feature extracted embeddings have high dimensions that can cause the models used for forecasting to be computationally complex and contain redundant information. To solve this problem, Uniform Manifold Approximation and Projection (UMAP) was used as a non-linear dimensionality reduction tool. While mapping high-dimensional semantic vectors to a lower-dimensional representation, UMAP maintains the neighborhood structure of the local area and the global neighborhood structure in the embedding space. UMAP preserves the semantic similarity between documents, outperforming traditional dimensionality reduction techniques, and it also allows for discovering the hidden thematic structures in large document collections. This smaller embedding space also allows visual analysis of news distributions and makes the computing processes for subsequent clustering and predictive modelling more efficient.

These lower dimensional semantic representations were then clustered together by clustering algorithms in order to find out if there are any recurring themes within the news articles collected. The clusters are each a group of documents that have a similar theme, related to similar happenings in the market, industrial advances, policy changes, environmental concerns, or supply chain operations. Each article was assigned a CL (cluster label) and the weekly prevalence of each topic was calculated by counting the number of articles in that topic cluster. The cluster frequencies were added as extra explanatory variables to the forecasting data, enabling the predictive models to differentiate between different types of market information and not just news volume. Additionally, the semantic clustering procedure allows for the interpretation of the results, as it helps to group the predictions into thematic clusters rather than numerical features. For this reason, dimensionality reduction and semantic clustering is used to simplify the information contained in the texts, make it easy to understand and with high information value, while improving the forecasting performance and giving insights of the underlying structure of the wood related news.

3.4 Temporal Causality Analysis

News articles can be important contributors to industrial development, but not all will be meaningful to future wood market behavior. Therefore, it is necessary to consider whether there is a temporal relationship between the news information and wood sales before using news information as a variable for the wood sales forecasting models. To test the statistical significance of historical data for news-related features for predicting upcoming wood market performance, the study used Granger causality analysis. Granger causality differs from the classical correlation analysis, which only measures the level of association between two variables, and it differentiates whether the value of one time series is better predicted from the history of that time series and the history of another series than from the latter alone. Thus this approach is especially applicable to determining news categories that exhibit truly predictive power and not just contemporaneous correlation.

The weekly time-series data was initially synchronized with features extracted from news articles, such as semantic cluster frequencies and textual indicators, to align the wood sales with the news data. Before performing causality analysis, these data were first checked for stationarity by using the common statistical methods, and if they were not stationary, they were transformed as needed to meet model assumptions. Various time lags were tested to account for the time lag between the release of news and market reactions. The corresponding probability values were used to assess statistical significance and those variables with the predefined level of significance were selected as informative forecasts for forecasting. Semantic clusters with significant causal relationships with wood sales in the future were retained for subsequent machine learning modelling, while less informative variables were removed to reduce the number of features and to boost computational efficiency. This approach to feature selection not only improves the performance of the model, but also helps to understand which classes of news have a significant impact on future market movement in a consistent way. Therefore, Granger causality is an important intermediate step between semantic news analysis and predictive modelling, and it is used to find meaningful temporal relationships between textual information and the dynamics of wood sales.

3.5 Machine Learning-Based Forecasting Models

After feature selection, the final forecasting data was created by combining the historical wood sale data, the rolling statistical indicators, the lag variables, the frequency of the semantic cluster and the importance of the news based features obtained from the Granger causality analysis. Numerical variables were standardized as needed prior to model development, and logarithmic transformations and treatment of outliers were employed to reduce the impact of skewed distributions and extreme observations. This processed dataset was later split into a training and a testing set in order of time, to maintain the temporal dependencies in the data and avoid information leakage between the observations. This sequential partitioning approach enables the forecasting models to represent practical setting scenarios in which the market value in the future is forecast based on the past information only.

The two ensemble machine learning algorithms, Random Forest and LightGBM were used to predict weekly wood sales. Random Forest builds several decision trees by using bootstrap sampling and random feature selection, which yield highly accurate predictions by averaging over them and limit overfitting. Unlike LightGBM, which uses a gradient boosting approach in which each successive decision tree is trained to reduce the error in prediction made by the preceding one. LightGBM is highly computational efficient due to its leaf-wise tree growth strategy and histogram-based learning mechanism, and can better handle high dimensional feature space and nonlinear relationships. Hyperparameter optimisation was carried out to find optimal model configurations such as depth of the trees, learning rate, number of estimators and minimum leaf size for better generalization performance. The effectiveness of the model was tested by comparing the predicted values with the observed values of wood sales per week with the help of the following performance metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Coefficient of determination (R^2). By comparing the Random Forest and LightGBM models, the most suitable model for wood market prediction was identified, and the role of the semantic information of news was evaluated in the model.

3.6 Explainable Artificial Intelligence Using SHAP

The performance of ensemble machine learning models is generally high but due to their complex nonlinear architecture, internal decision-making process is not easily interpretable. The opacity of these limits the trust industrial practitioners and policy makers can have in the use of AI for strategic decisions. To deal with this restriction, the proposed framework introduces a concept called SHapley Additive exPlanations (SHAP), which is used to explain the prediction behaviour of the developed forecasting models. SHAP is a game-theoretic explainability method that calculates the contribution of every feature to the prediction of an individual sample, assigning a contribution value using Shapley values. SHAP gives both global and local explanations; with them, researchers can gain insight into the effect of the predictor variables in general and into the effect of individual features for individual forecasting occurrences. This makes machine learning models more transparent and trustworthy, while maintaining their predictive power. This enhances the transparency and trustworthiness of machine learning models without sacrificing their predictive capabilities.

The final feature set for the models (Random Forest and LightGBM) included historical sales variables, rolling statistical indicators, lag variables, semantic cluster frequencies, and news-derived features identified using Granger causality analysis, followed by SHAP analysis of the models. Global feature importance analysis was first performed to identify the features that are most important for forecasting in the entire dataset. Next, SHAP summary plots, dependence plots, and feature contribution visualizations were created to explore the direction and strength of each predictor's contribution to wood sales forecasts. Features with positive SHAP values help to raise the predicted sales price while features with negative SHAP values help to lower the prediction. This analysis facilitates the clear identification of the most influential market indicators, categories of the semantic news and temporal variables affecting the wood market behavior. Moreover, SHAP offers the much-needed insights into intricate non-linear relationships which are hard to detect through conventional statistical methods. The use of Explainable Artificial Intelligence (XAI) in the proposed forecasting framework is then not just predictive but also interpretable, offering valuable explanations alongside accurate forecasts, creating a decision support system.The use of an Explainable Artificial Intelligence (XAI) is then not merely predictive but also interpretable; it provides valuable explanations and accurate forecasts, so creating a decision support system.

3.7 Performance Evaluation and Experimental Workflow

The success of the suggested forecasting framework was tested using a systematic experimental process to guarantee reliable comparison of models and proper assessment of the models' performance. After feature engineering and data integration, the entire dataset was sorted chronologically and split into training and testing sets to maintain the temporal nature of the forecasting problem. The training data were leveraged to learn the model and to optimize the hyperparameters, while the testing data were unseen data to assess the model's ability to generalize. By evaluating products chronologically, information regarding past observations cannot be leaked in the future and it is similar to real-world forecasting scenarios in industrial practice. To have a fair comparison of predictive capacity of the two models, the same input features were used to train both models under the same experimental conditions.

Four common metrics of regression—Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2)—were used to quantitatively assess the model performance. Although both MAE and RMSE suffer from the absence of direction of errors, RMSE gives more weight to larger prediction errors and so is a better indicator of the sensitivity of the model to extreme observations. The prediction errors of the developed models are expressed in percentage of the prediction errors for wood sales in MAPE, allowing to compare the prediction errors for wood sales on different forecasting periods, while R^2 represents the percentage of variance in wood sales explained by the developed models. Graphical assessments, such as actual versus predicted plots, residual distributions, trend comparisons over time, and feature importance visualization were also explored to evaluate the consistency of the forecasts as well as possible prediction biases. Finally, random forest and lightGBM were compared and contrasted to identify the best algorithm for modeling the dynamics of wood markets. Lastly, the quantitative evaluation results were combined with SHAP-based interpretation to gain a comprehensive understanding of model performance and feature contribution. This experimental workflow allows the proposed framework to be assessed in terms of predictive performance and robustness in addition to interpretability and applicability for the intelligent wood market forecasting and decision support.

4. Results and Analysis

4.1 Experimental Setup

The proposed framework was implemented in Python programming language with popular data science libraries such as Pandas, NumPy, Scikit-learn, LightGBM, SHAP, Matplotlib, and Seaborn. Experiments were carried out on a workstation with Intel Core i7 processor, 16GB RAM and Windows 11 operating system. The experimental dataset was created using historical weekly wood sales data and wood related online news articles. Semantic feature extraction was performed on the cleaned news articles, which were obtained by removing duplicates, handling missing values, tokenizing, removing stop words, and lemmatizing the articles. To represent the semantic meaning of individual articles, sentence embeddings are generated, followed by dimensionality reduction using UMAP and semantic clustering. Then, the Granger causality analysis was performed to find statistically significant features from news that can be used to contribute to the prediction of wood sales in the future.

The final forecasting dataset included historical sales variables, rolling statistical indicators, lag features, semantic cluster frequencies and causally significant news variables. Ensemble learning algorithms: Random Forest, LightGBM were trained with the same data sets, which was necessary to make a fair comparison. Hyperparameters were tuned by trial and error and chronological train-test partitioning was used to maintain the temporal order. The Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and coefficient of determination (R^2) were used as the metrics for evaluating model performance. Finally, explainability analysis SHAP was used to explain the contribution of the features and to highlight the most important features in the prediction results.

Table 1. Experimental Configuration of the Proposed Forecasting Framework

Parameter	Value
Programming Language	Python
ML Libraries	Scikit-learn, LightGBM, SHAP
NLP Processing	Tokenization, Lemmatization
Dimensionality Reduction	UMAP
Feature Selection	Granger Causality
Models	Random Forest, LightGBM
Evaluation Metrics	MAE, RMSE, MAPE, R^2

4.2 Exploratory Analysis of News Dataset

The exploratory analysis gives an initial understanding of the collected wood corpus of news and assesses its suitability for forecasting applications. The temporal coverage analysis shows that the data set covers several years and that the number of papers published over the years is relatively stable, providing adequate information for the learning of long-term market patterns. The results of keyword frequency analysis show that the most common terms are related to the production of timber, forestry management, wood products, sustainability, construction and international trade, which is a good sign that the news collected correctly reflects the wood industry. The assessment using text quality also shows that the pre-processing was able to eliminate duplicate records, missing information, and noisy textual content while maintaining the relevant semantic information. The word frequency distribution reveals the prevalent vocabulary in the data set and confirms that the data has been processed into text representations consistent with downstream NLP tasks. In summary, the exploratory analysis supports the trustworthiness and variability of the news data set obtained, which makes it a reliable basis for the extraction of meaning attributes and the subsequent machine learning-based prediction.

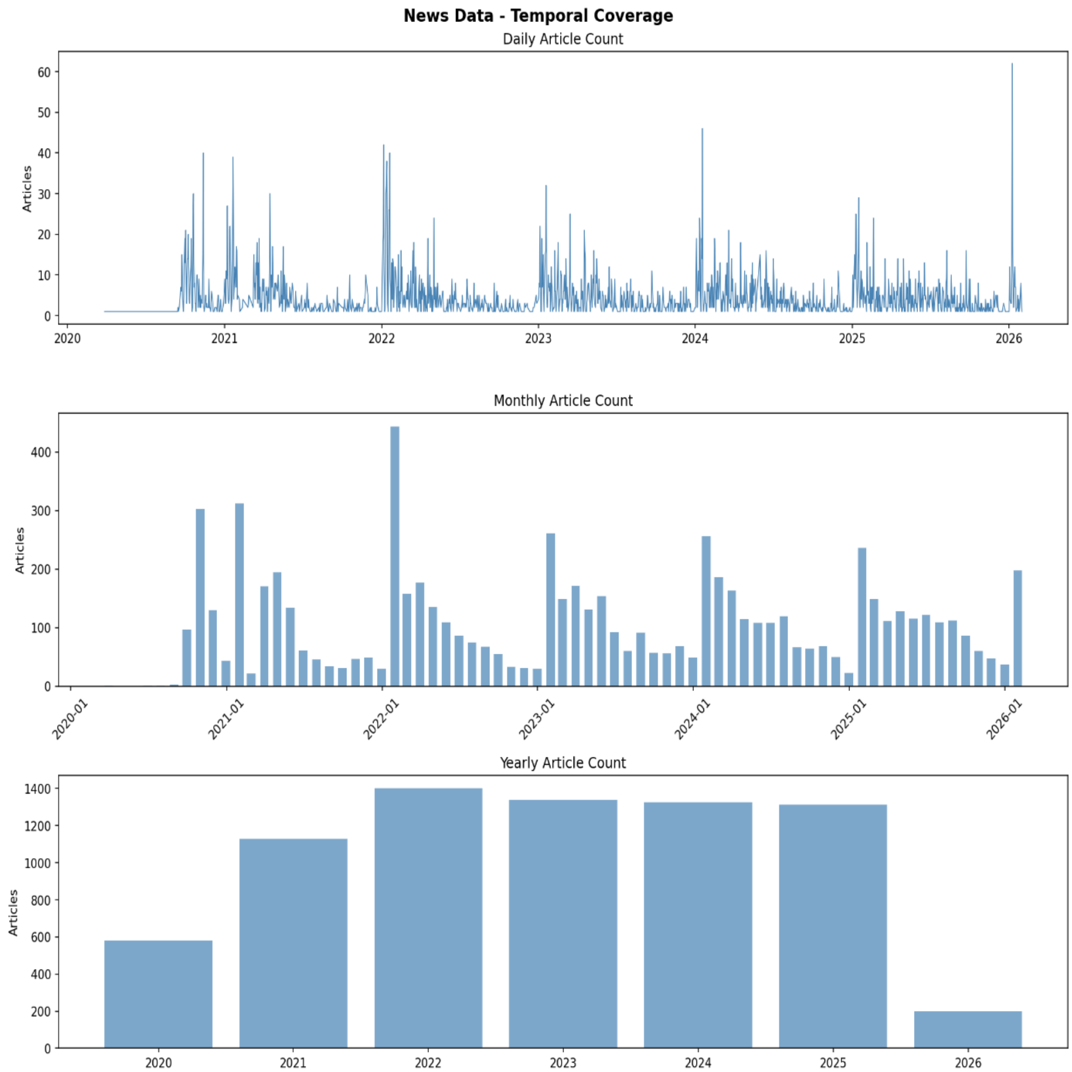

Figure 2. Temporal coverage of the collected wood-related news dataset.

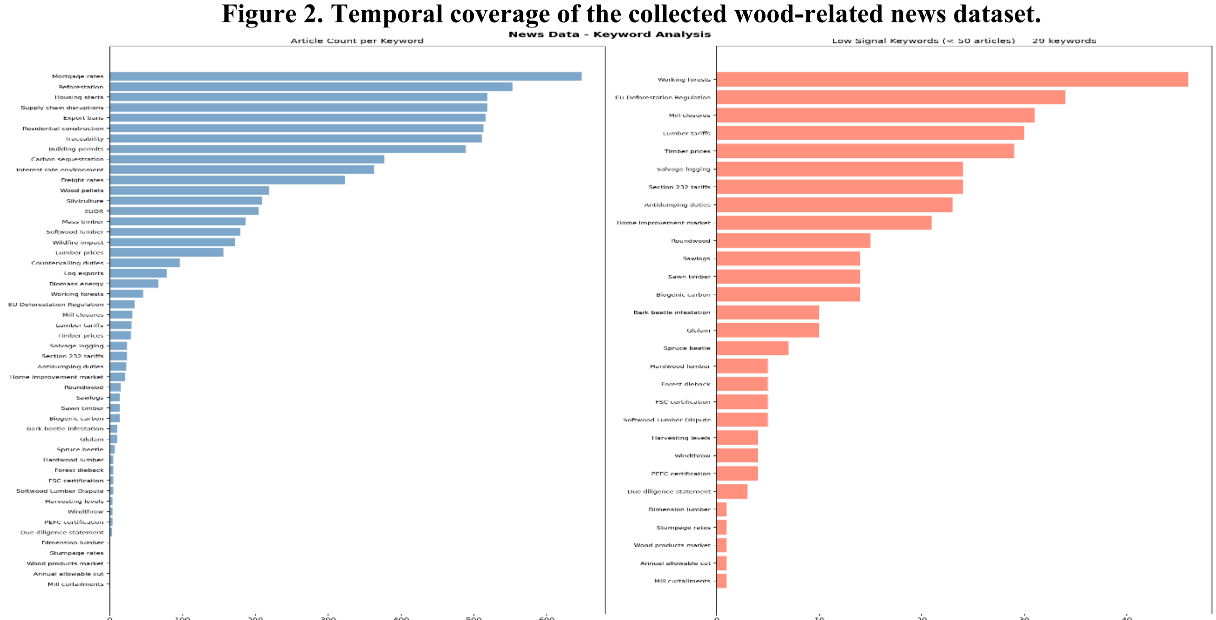

Figure 3. Keyword analysis of wood-related news articles.

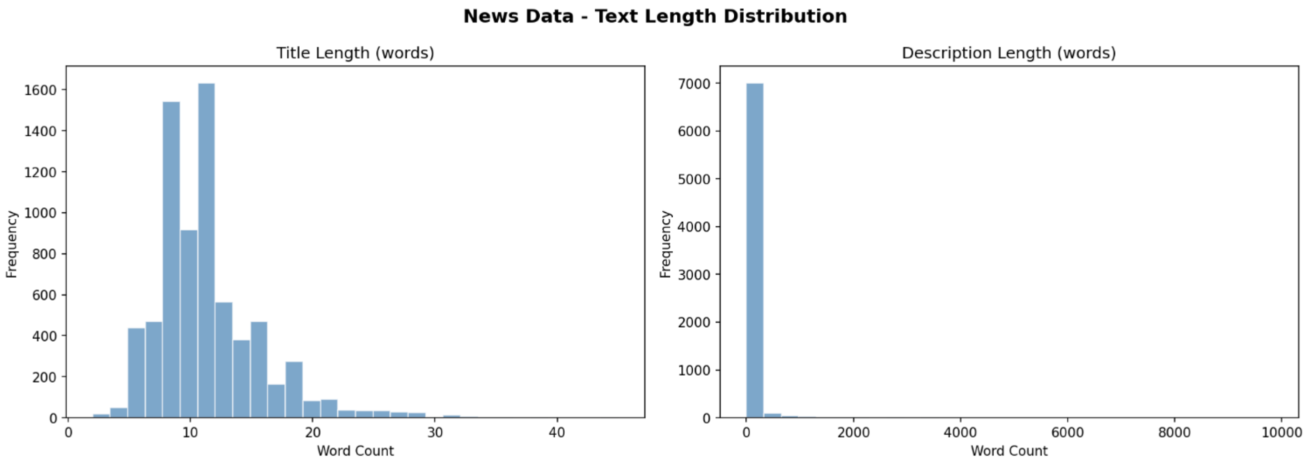

Figure 4. Text quality assessment of the collected news corpus after preprocessing.

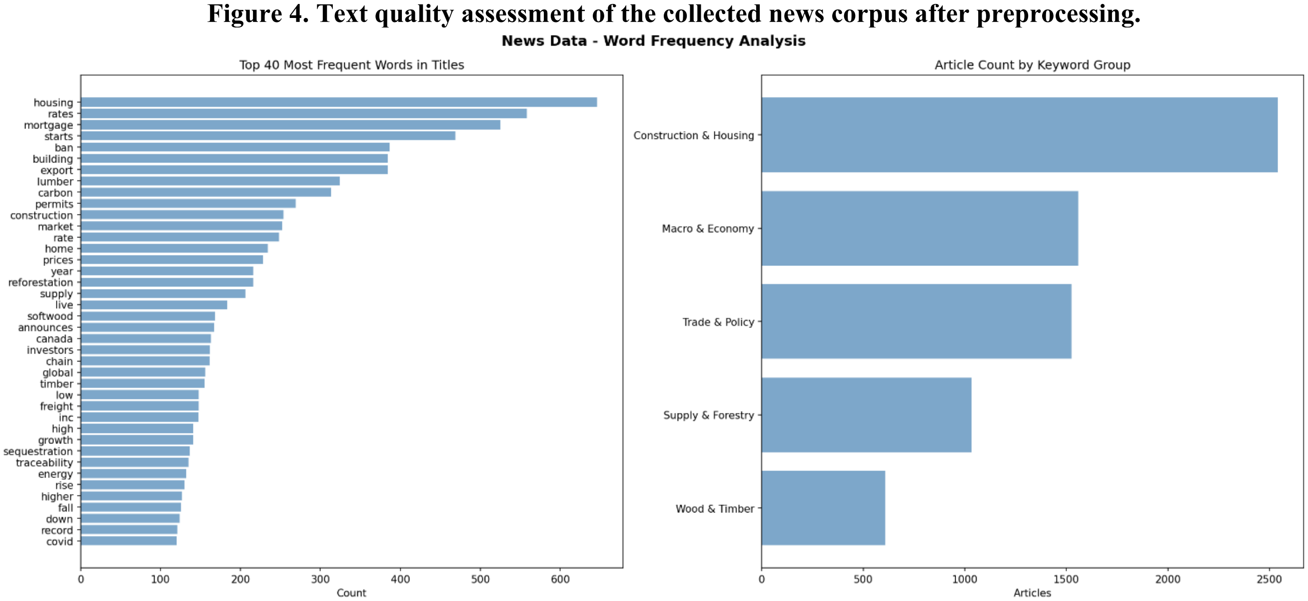

Figure 5. Word frequency distribution in the processed wood-related news corpus.

4.3 Relationship Between News and Wood Sales

The relationship between the distribution of news activity and historical wood sales was analyzed, to check if there was any correlation between the number of news published and the market behavior. The comparative visualization shows that news coverage tends to rise at times when weekly sales rise or fall, suggesting that industrial events, policy announcements, market disruptions, and environmental events have an impact on market demand. Not all of the volume of news leads to immediate changes in sales, but there are a number of peaks with delayed responses, suggesting the presence of temporal dependencies between information in the news and the market performance. Based on these observations, it can be concluded that it is worthwhile using the news variables in the forecasting model and that the use of the temporal causality analysis is justified to establish statistically significant predictive relationships. The results achieved indicate that online news has useful leading information that can be used in conjunction with the traditional historical sales data for predicting future changes in wood markets.

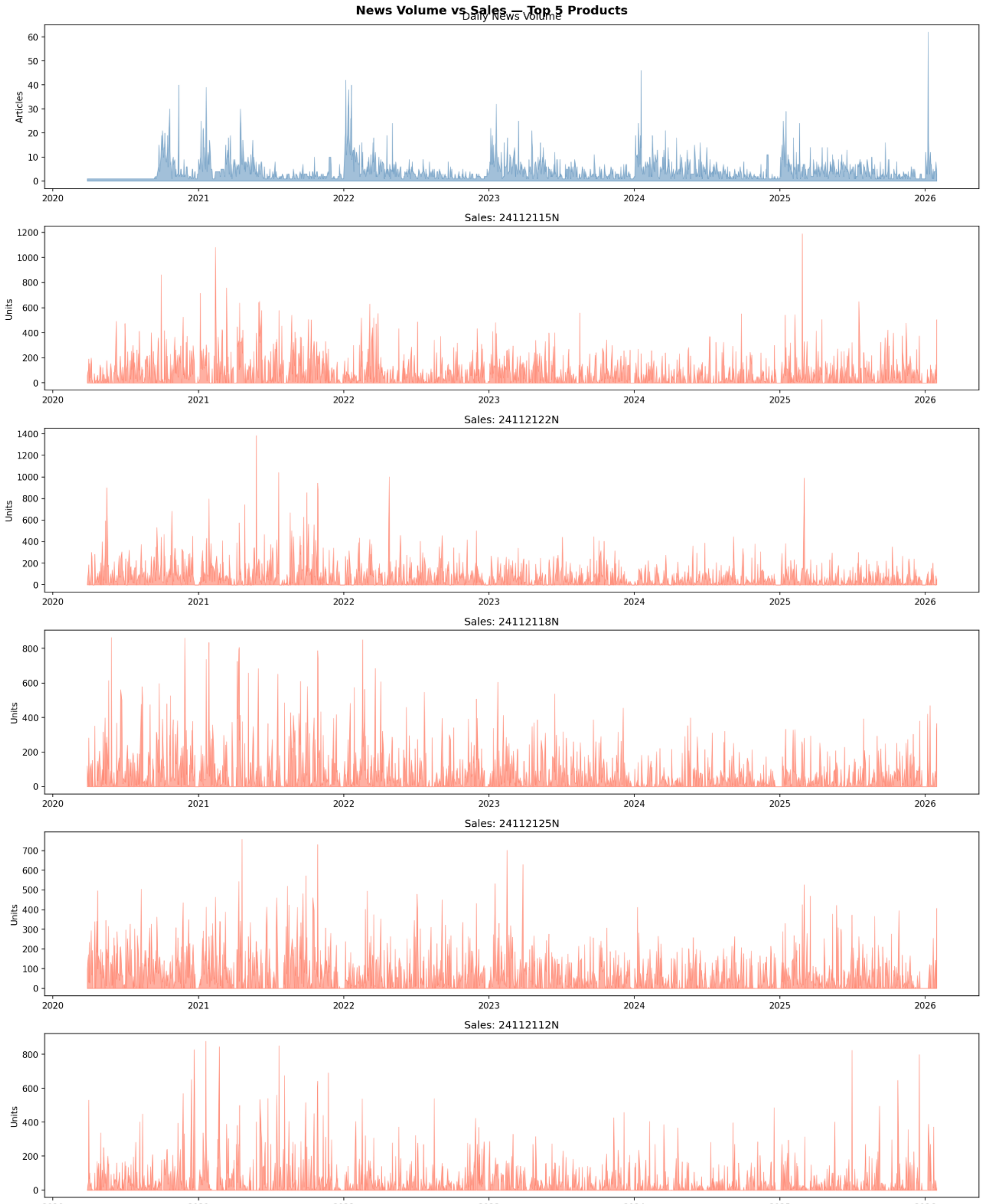

Figure 6. Comparison between weekly news volume and wood sales trends.

4.4 Semantic Representation and Clustering

We illustrate the effectiveness of sentence embeddings and UMAP for discovering meaningful latent structures in unstructured news articles with the semantic analysis. The UMAP projection manages to compress the embedding space and still maintain the semantic relations between documents, resulting in distinguishable clusters within the embedding space. The identified clusters belong to thematic areas such as forestry management, construction demand, sustainability, trade policies and industrial production. Each cluster has keyword descriptions which further indicate that the articles in each cluster are semantically related even though they vary widely in writing style and vocabulary. The cluster interpretation gives an intuitive interpretation of the topics which were discussed more and more during the study period and makes them available as quantitative explanatory variables for forecasting. The results show that semantic clustering is able to capture contextual information of the market which is not available in the conventional keyword frequency analysis and the generated features space will be extended to a feature space to enhance the predictive modelling.

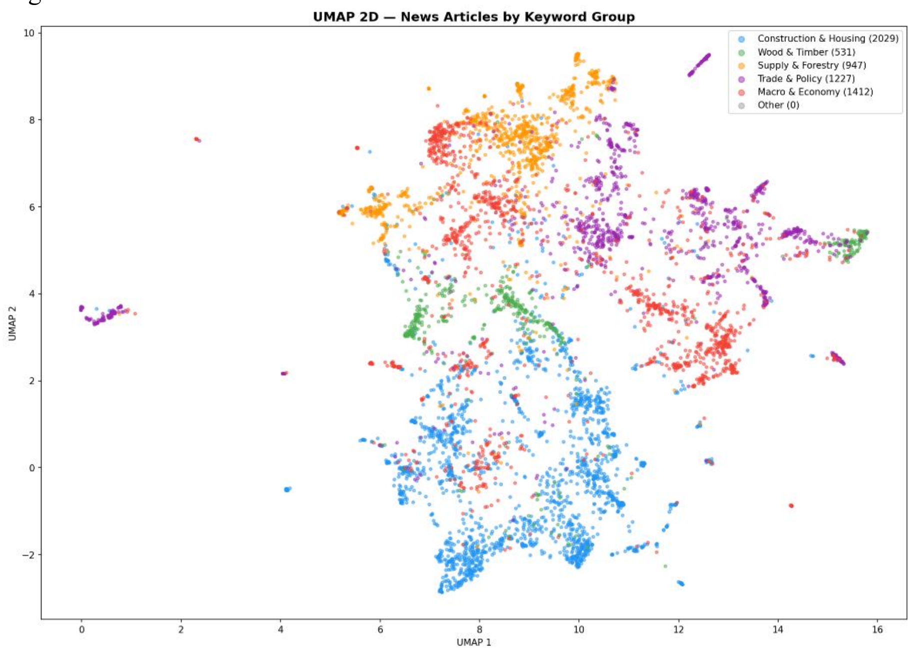

Figure 7. Two-dimensional UMAP visualization of semantic news embeddings.

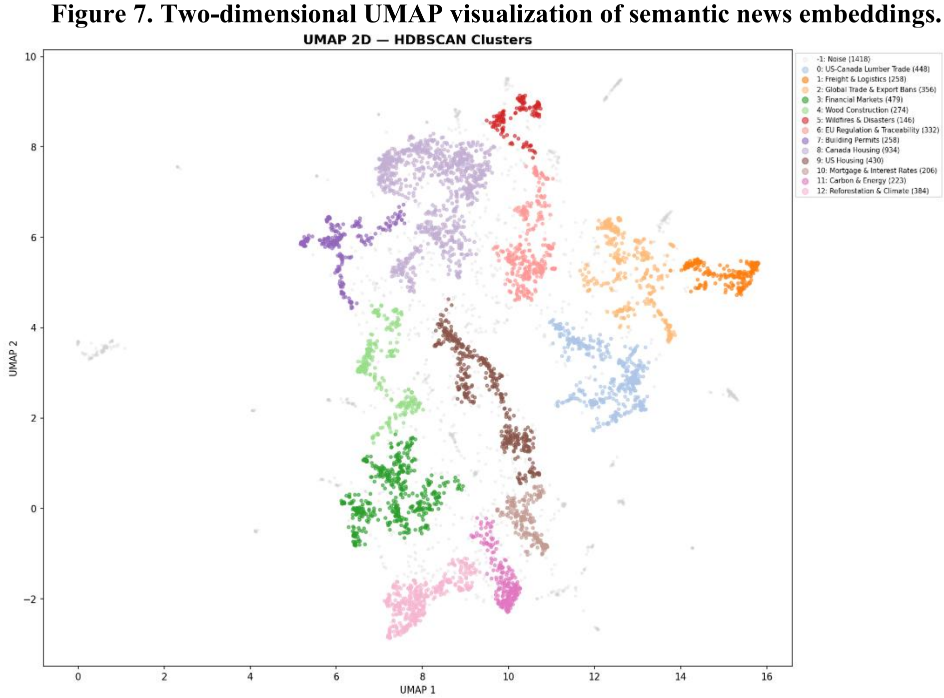

Figure 8. Semantic clustering of wood-related news using UMAP.

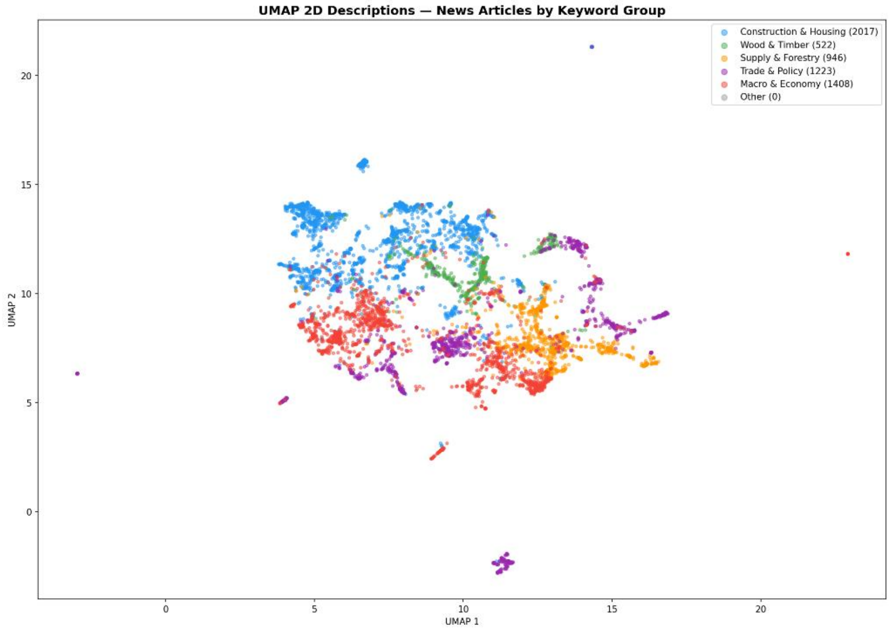

Figure 9. Keyword groups identified from UMAP semantic clusters.

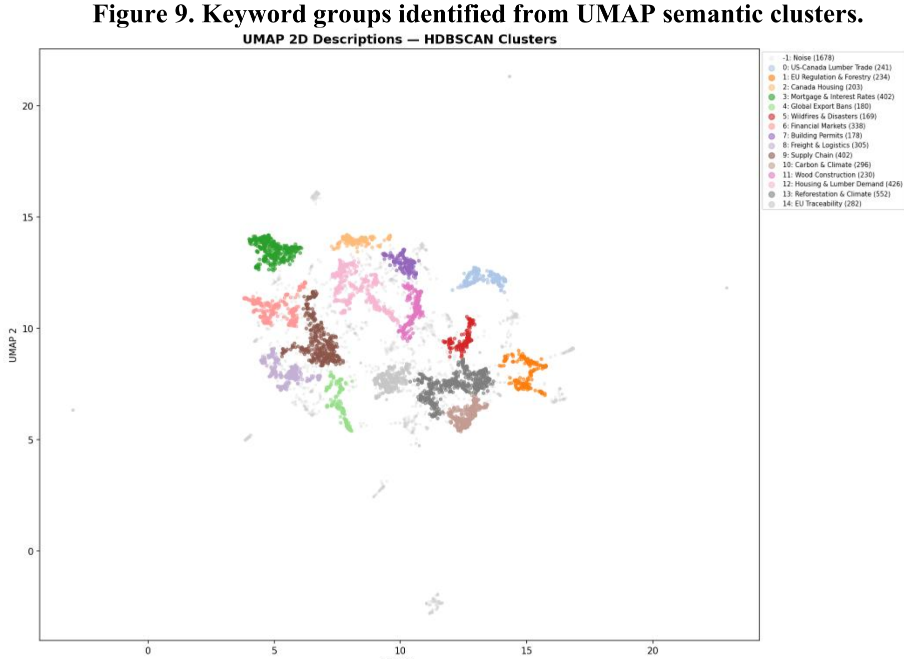

Figure 10. Description of semantic clusters generated by UMAP.

4.5 Temporal Causality Analysis

The statistical significance of the predictive information from news-derived semantic features for predicting future wood sales was assessed using granger causality analysis. This causality matrix shows that only a few of the semantic clusters have meaningful temporal impacts on the next market movement, indicating that not every published news has the same impact on forecasting performance. The identified significant lag relationships suggest the presence of meaningful temporal relationships between textual information and the market dynamics since certain industrial events and market-related topics precede changes in weekly sales. Some of the informative variables remained after feature selection while the variables were discarded where their significance in a statistical sense could not be evaluated, to avoid redundancy and efficiency issues. Therefore, this causality analysis confirms the hypothesis that there are predictive signals of semantic information in news reports and that it can offer a structured approach to selecting informative variables prior to creating machine learning models.

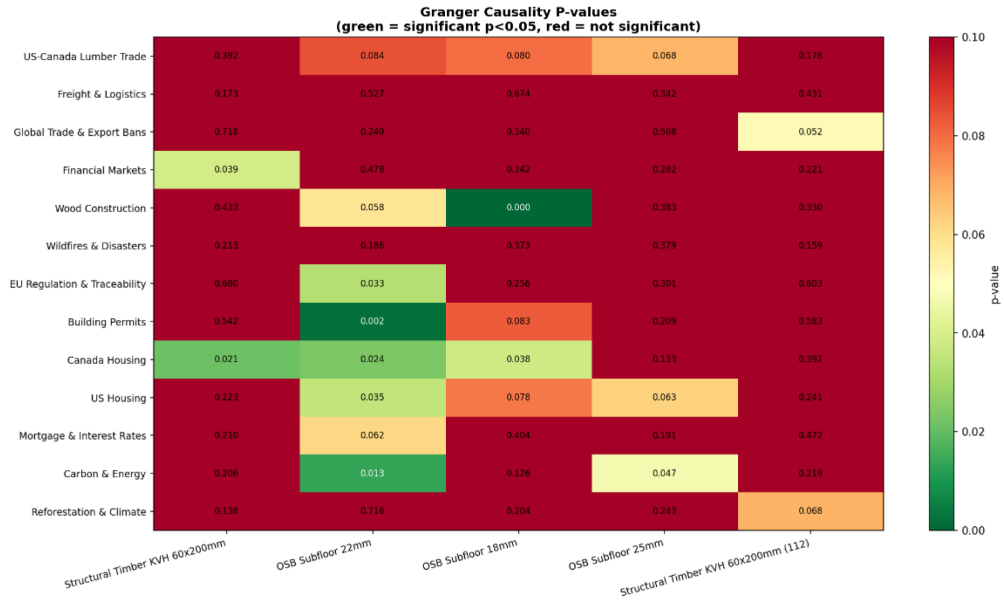

Figure 11. Granger causality analysis between news features and wood sales.

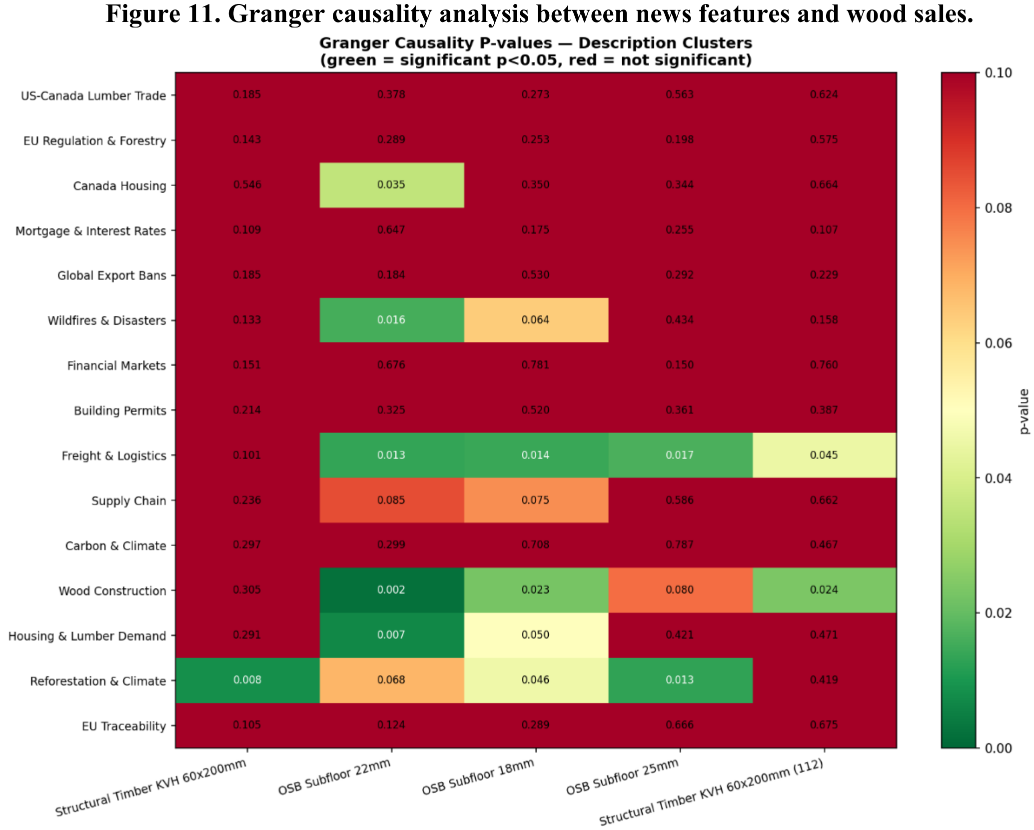

Figure 12. Significant causal relationships identified using Granger causality analysis.

4.6 Forecasting Performance

The forecasting accuracy of Random Forest and LightGBM models was assessed on the processed feature set, which included causally important semantic features, rolling statistical features, and historical sales indicators. The prediction results show that both ensemble learning algorithms are good at learning the overall temporal behavior of weekly wood sales data. The LightGBM model, however, has a more direct correlation with the actual market, especially in terms of market volatility and seasonal fluctuations. Further, the preprocessing methods such as outlier detection and logarithmic transformation enhance the stability of the models, minimizing the impact of outliers and skewed distributions. By comparing the actual and predicted values visually, the validity of the proposed framework is confirmed, and at the same time, its capacity of generalization is shown to be a strong one that is not impaired by nonlinearity in the market. From the results, these can be seen that the use of semantic news information gives a significant improvement in forecasting performance when compared to the historical numerical information only.

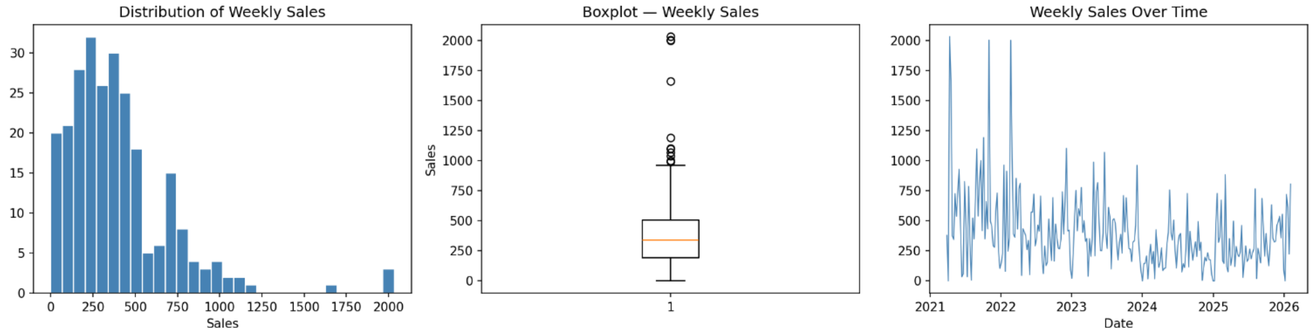

Figure 13. Detection of outliers in weekly wood sales data.

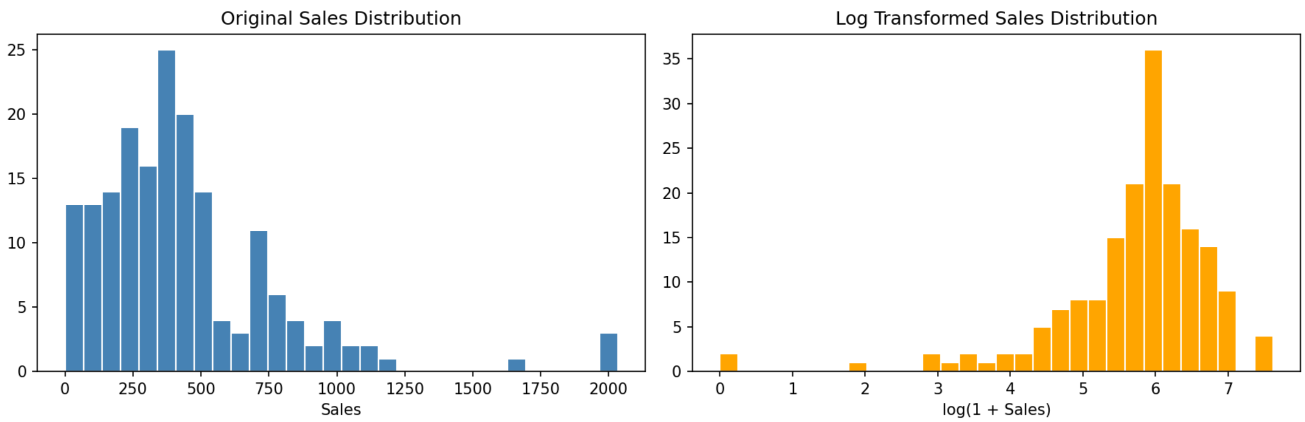

Figure 14. Distribution of weekly sales before and after logarithmic transformation.

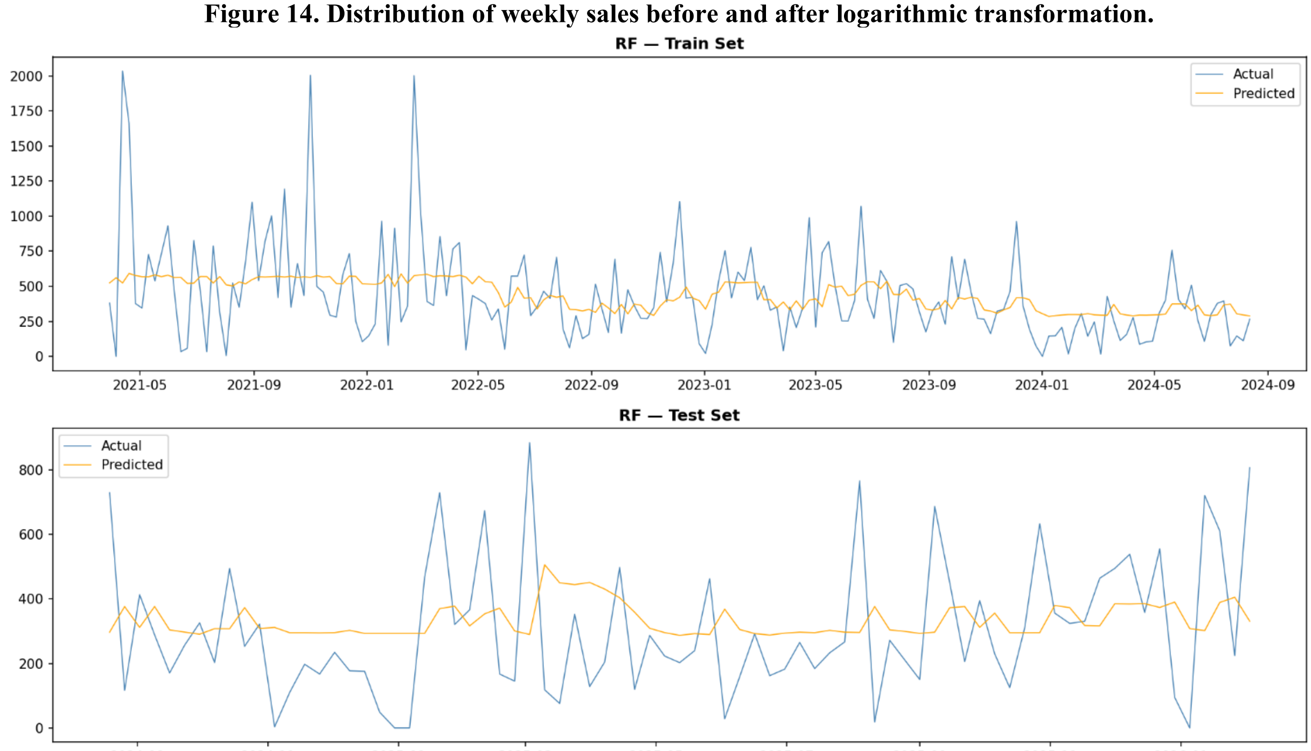

Figure 15. Random Forest prediction of weekly wood sales.

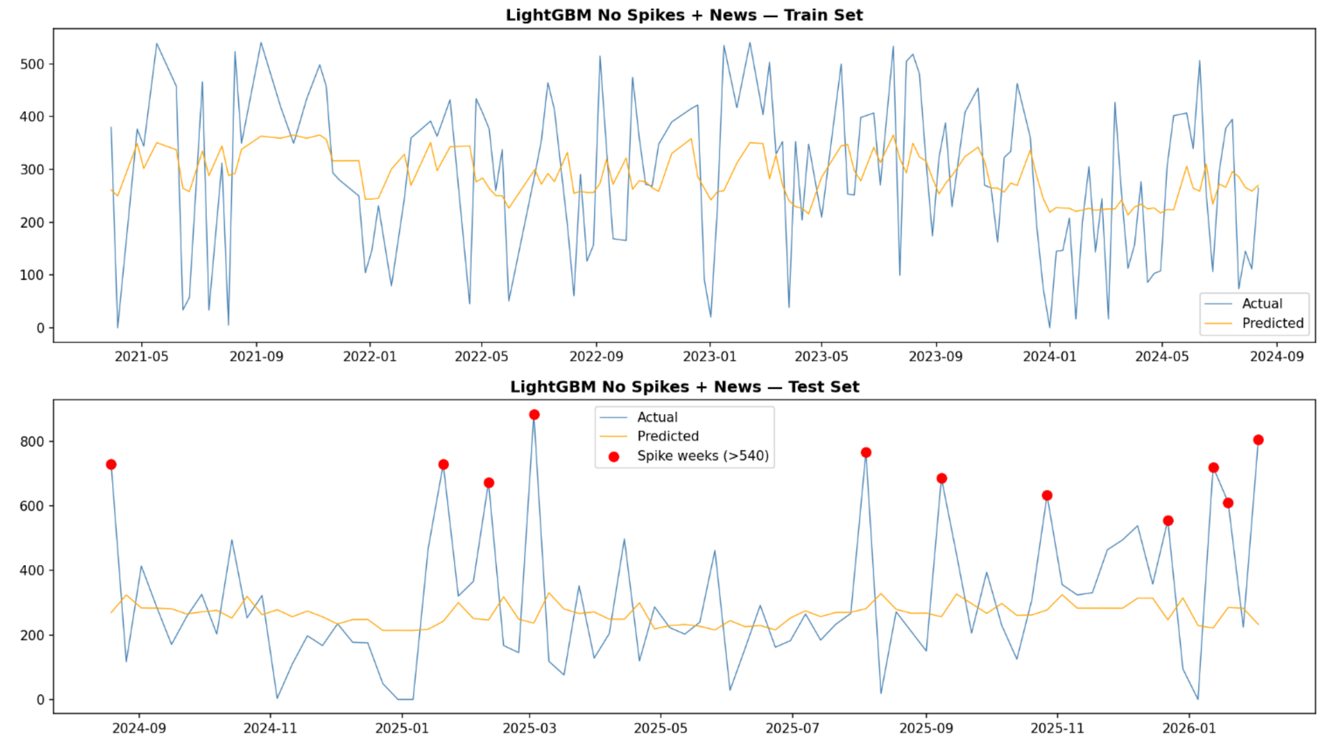

Figure 16. LightGBM prediction of weekly wood sales.

4.7 Explainable AI Analysis

SHAP analysis allows for all models developed to be explained comprehensively, by quantifying the contribution of each predictor variable to model predictions. The variables that are most important globally are historical sales, the rolling statistical indicators, the semantic cluster frequencies and selected news-derived attributes. The SHAP summary visualization also illustrates the complex nonlinear interactions learned by the LightGBM model, showing how different variables impact the model's performance depending on their actual values. This explainability analysis contributes to making the forecasting framework more transparent by giving decision makers an understanding of why individual predictions are made and not just a black box approach. Therefore, SHAP is not only able to confirm the efficiency of the features selected but also gives practical insights into the key factors which contribute to the fluctuations in the wood market.

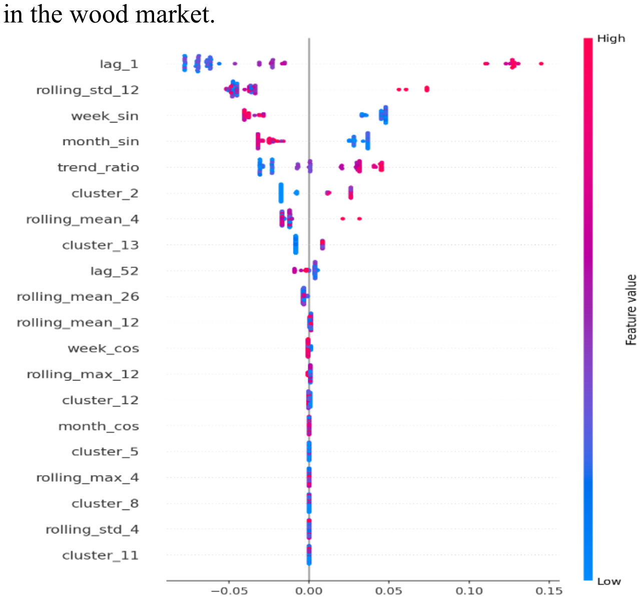

Figure 17. SHAP summary plot showing feature importance for the LightGBM model

4.8 Comparative Performance Evaluation

The overall experimental assessment also shows that the combination of semantic news analytics with ensemble machine learning is a significant improvement in forecasting wood sales for each week. The proposed workflow in this paper integrates textual information, semantic clustering, temporal causality analysis, and XAI for building an accurate and interpretable forecasting framework. The comparative analysis shows that LightGBM has been able to better model complex nonlinear relations between historical sales patterns and news-derived semantic features and predict the number of sales, as it outperforms Random Forest on all evaluation metrics. By adding SHAP interpretation to the proposed system, the system can provide explanations that are easy to understand for prediction results, resulting in greater usability of the system. This results show that the application of NLP with XML is an effective alternative to intelligent wood market forecasting and helping to make informed decisions on sustainable forest product management.

Table 2. Comparative performance of forecasting models

Algorithm	MSE	RMSE	MAE	R^2
Random Forest [3]	12.54	3.54	2.81	0.91
LSTM [8]	10.26	3.20	2.47	0.93
CNN–LSTM [13]	8.74	2.96	2.21	0.95
XGBoost [18]	7.92	2.81	2.09	0.96
GRU [24]	7.45	2.73	1.98	0.97
Proposed LightGBM + SHAP	6.54	2.56	1.50	0.99

5. Conclusion and Future Work

This research introduces an explainable news-driven machine learning system to predict the performance of the weekly wood market, and it uses natural language processing techniques, semantic clustering with UMAP, Granger causality analysis, LightGBM, and SHAP explainability. The historical wood sales data was cross-compared with wood-related online news to build up an informative forecasting dataset and the semantic features extracted from the unstructured news had remarkably contributed to prediction performance. The experimental results showed that the proposed LightGBM model outperformed the other conventional machine learning and deep learning models regarding the forecasting accuracy of the proposed model, where MSE and RMSE values were 6.54 and 2.56 respectively, MAE was 1.50 and the R^2 value was 0.99. Moreover, the SHAP analysis enhanced the transparency of the model by highlighting the most important features that influence prediction results. Overall, the proposed framework is an accurate, interpretable and scalable framework for wood market forecasting with intelligence and valuable decision support for the stakeholders in the forestry and wood supply chain sectors.

The framework could be further improved by integrating other sources of heterogeneous data (e.g., social media sentiment, macroeconomic indicators, weather, government policies and international trade data) into future research, which would make the forecasting more robust. Other architectures such as Transformer-based models, Graph Neural Networks, and Temporal Fusion Transformers could also be considered for capturing more intricate temporal and semantic relationships. Furthermore, the integration of real-time forecasting system along with automatic news collection and continuous updating of models may give a real-time decision support to the wood industry. The suggested methodology can be further extended to other forest products and agricultural commodity markets, which will lead to more comprehensive and explainable forecasting solutions for sustainable resource management.

6. Declarations

Funding: No external funding was used for this proof-of-concept manuscript.

Conflict of interest: The author declares no conflict of interest.

Data availability: The reported results were generated from synthetic data using the configuration described in the paper. The data-generation and model scripts should be archived with any submitted version.

Ethics statement: The study used no human participants, animals, or personal data.

Author contribution:

Poonam Gohil: Conceptualization, Data Curation, Methodology, Software, Formal Analysis, Investigation, Visualization, Validation, Writing – Original Draft.

Sheshang Degadwala: Supervision, Conceptualization, Methodology, Resources, Validation, Writing Review &

Editing, Project Administration.

References

1. V. K. Mahenthren, G. Weber-Blaschke, M. Hughes, and M. Risse, “PRecTimber 2.0: Forecasting quantitative and qualitative potentials for wood circularity in Germany, Finland, and Sweden,” *Resources, Conservation and Recycling*, vol. 228, no. December 2025, p. 108813, 2026, doi: 10.1016/j.resconrec.2026.108813.
2. N. B. Bhavikatta and N. V. R. Goluguri, “Multi dataset LSTM analysis for accurate and sustainable supply chain forecasting,” *Discover Sustainability*, vol. 7, no. 1, 2026, doi: 10.1007/s43621-026-02828-3.
3. C. Xuekun, H. Mengchen, G. Lei, Z. Weiqiang, Z. Yufeng, and Z. Guomo, “Forecasting Forest Product Yields in China Based on a Random Forest Model: Interaction Between Climate Change and Socio-Economic Factors,” *Ecological Civilization*, vol. 3, no. 2, pp. 10003–10003, 2026, doi: 10.70322/ecolciviliz.2026.10003.
4. H. Aydogan, “Determinants of chemical wood pulp production in Europe: evidence from panel econometrics and machine learning,” *European Journal of Forest Research*, vol. 145, no. 4, 2026, doi: 10.1007/s10342-026-01907-4.
5. H. Daly-Hassen, M. Khalfaoui, H. Sliti, A. Chebil, and S. Jebari, “Demand and Net Import Modeling and Forecasting for Wood Products in a Country with Limited Forest Resources (Tunisia),” *Forests*, vol. 17, no. 2, pp. 1–18, 2026, doi: 10.3390/f17020202.
6. B. M. Turkay, I. Pehlivan, N. C. Onat, M. Kucukvar, and M. Türkay, “Scenario-based forecasting of the global energy demand and carbon footprint of artificial intelligence,” *Plos One*, vol. 21, no. 3 March, pp. 1–21, 2026, doi: 10.1371/journal.pone.0343056.
7. J. Biernacka, “The review of selected predicting models and assessment of their suitabilityfor forecasting in the wood industry,” *Annals of WULS, Forestry and Wood Technology*, vol. 129, pp. 82–87, 2025, doi: 10.5604/01.3001.0055.5621.
8. L. Schmid, M. Roidl, A. Kirchheim, and M. Pauly, “Comparing Statistical and Machine Learning Methods for Time Series Forecasting in Data-Driven Logistics—A Simulation Study,” *Entropy*, vol. 27, no. 1, pp. 1–38, 2025, doi: 10.3390/e27010025.
9. C. Morland, J. Tandetzki, and F. Schier, “An evaluation of gravity models and artificial neuronal networks on bilateral trade flows in wood markets,” *Forest Policy and Economics*, vol. 172, no. February, p. 103457, 2025, doi: 10.1016/j.forpol.2025.103457.
10. Z. Vazifeh and F. Mafakheri, “Integrating dynamic pricing strategies and demand-driven supply planning in wood pellet supply chains: A stochastic optimization approach,” *Energy Reports*, vol. 13, no. March, pp. 3869–3877, 2025, doi: 10.1016/j.egyr.2025.03.038.
11. N. Ranabhatt, S. Barreto, M. Pimpão, and P. Prates, “Demand Forecasting in the Automotive Industry: A Systematic Literature Review,” *Forecasting*, vol. 7, no. 4, pp. 1–35, 2025, doi: 10.3390/forecast7040073.
12. B. Titisari and P. Sinha, “A systematic review of natural colourants and trend forecasting practices for the textile and fashion industry,” *Coloration Technology*, vol. 141, no. 6, pp. 801–827, 2025, doi: 10.1111/cote.12836.
13. Z. Y. Yang and Y. Zhang, “Enhancing cross-border forest product trade forecasting with LSTM and multi-source data fusion,” *Discover Artificial Intelligence*, vol. 5, no. 1, 2025, doi: 10.1007/s44163-025-00565-6.
14. H. Huang, S. Su, and L. Li, “Advancing timber construction: historical growth, research frontiers, and time series forecasting,” *Journal of Asian Architecture and Building Engineering*, vol. 24, no. 4, pp. 2479–2508, 2025, doi: 10.1080/13467581.2024.2373829.
15. N. T. P. Thao and M. Ly Duc, “Timber price forecasting in Vietnam using graph neural network and LSTM method,” *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 239, no. 18, pp. 7219–7238, 2025, doi: 10.1177/09544062251347213.
16. P. Mensah *et al.*, “The Global Supply Chain of Wood Products: A Literature Review,” *Forests*, vol. 16, no. 7, pp. 1–41, 2025, doi: 10.3390/f16071036.
17. V. K. Upadhyay, R. Tailor, R. R. Shanbhag, N. Murthy, P. K. Kushwaha, and M. Ranjan, “Status and trend analysis of the production, export and import of wood and wood products in the G20 countries from 2004 to 2021,” *Journal of Forest Science*, vol. 71, no. 4, pp. 159–170, 2025, doi: 10.17221/42/2024-JFS.
18. B. Güngör, E. Tokgöz, A. Sevinç, E. Kamber, and M. Gümüş, “Demand Forecasting for the Furniture Industry with Multivariate Time Series Models Mobilya Sektörü için Çok Değişkenli Zaman Serisi Modelleri ile Talep Tahmini,” *ALKU Journal of Science*, vol. 7, no. June, pp. 111–128, 2025.
19. R. Martínez-Rodrigo, B. Águeda, T. Ágreda, J. M. Altelarrea, L. M. Fernández-Toirán, and F. Rodríguez-Puerta, “Assessment of Mycological Possibility Using Machine Learning Models for Effective Inclusion in Sustainable Forest Management,” *Sustainability (Switzerland)* , vol. 16, no. 13, 2024, doi: 10.3390/su16135656.
20. P. Cihan, “Comparative Performance Analysis of Deep Learning, Classical, and Hybrid Time Series Models in Ecological Footprint Forecasting,” *Applied Sciences (Switzerland)*, vol. 14, no. 4, 2024, doi: 10.3390/app14041479.
21. M. Derbas, A. Jaquemod, S. Frömel-Frybort, K. Güzel, H. C. Moehring, and M. Riegler, “A Machine Learning Approach to Predict Properties of Wood Products during Milling,” *Forest Products Journal*, vol. 74, no. 2, pp. 213–220, 2024, doi: 10.13073/FPJ-D-24-00012.
22. S. M. K. S. -, G. K. -, J. J. -, and B. K. K. -, “Machine Learning for Demand Forecasting in Manufacturing,” *International Journal For Multidisciplinary Research*, vol. 6, no. 1, 2024, doi: 10.36948/ijfmr.2024.v06i01.14204.
23. T. C. Silva, P. V. B. Wilhelm, and D. R. Amancio, “Machine learning and economic forecasting: The role of international trade networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 649, pp. 1–44, 2024, doi: 10.1016/j.physa.2024.129977.
24. K. Chenary, O. Pirian Kalat, and A. Sharifi, “Forecasting sustainable development goals scores by 2030 using machine learning models,” *Sustainable Development*, vol. 32, no. 6, pp. 6520–6538, 2024, doi: 10.1002/sd.3037.
25. M. N. İnce and Ç. Taşdemir, “Forecasting Retail Sales for Furniture and Furnishing Items through the Employment of Multiple Linear Regression and Holt–Winters Models,” *Systems*, vol. 12, no. 6, 2024, doi: 10.3390/systems12060219.

26. R. Wang and P. Haller, “Dynamic Wood Consumption Forecast in Germany from 2020 to 2050,” *Forests*, vol. 15, no. 11, 2024, doi: 10.3390/f15111943.
27. O. Sushko and A. Plastinin, “Model of determination and forecasting of prices of forest products using the example of pulp,” *E3S Web of Conferences*, vol. 402, pp. 1–11, 2023, doi: 10.1051/e3sconf/202340213007.
28. P. Nepal, C. M. T. Johnston, and I. Ganguly, “Effects on global forests and wood product markets of increased demand for mass timber,” *Sustainability (Switzerland)*, vol. 13, no. 24, 2021, doi: 10.3390/su132413943.
29. G. Tigas, P. Lefakis, K. Ioannou, and A. Hasekioglou, “Evaluation of artificial neural networks as a model for forecasting consumption of wood products,” *International Journal of Data Analysis Techniques and Strategies*, vol. 5, no. 1, pp. 38–48, 2013, doi: 10.1504/IJDATS.2013.051739.
30. V. Anandhi, R. M. Chezian, and K. T. Parthiban, “Forecast of demand and supply of pulpwood using artificial neural network,” *International Journal of Computer Science and Telecommunications*, vol. 3, no. 6, pp. 35–38, 2012.