# Effective Cement Demand Forecasting using Deep Learning Technology: A Data-Driven Approach for Optimal Demand Forecasting

Sarita Nandal[1], Samyak Sabannawar[2], Akhil Pawan[3], Purvaja Fursule[4], Bharani Kumar Depuru[5]
[1,2]Research Associate, [3]Mentor, Research and Development, [4]Team Leader, Research and Development, [5]Director,
Innodatatics, Hyderabad, India

*Corresponding Author: Bharani Kumar Depuru
OCR ID: 0009-0003-4338-8914

**Abstract:- This research addresses the critical challenge of forecasting cement demand tailored to specific markets while concurrently optimizing distribution strategies to minimize transportation costs, reduce delivery times, and enhance inventory management. Leveraging a rich dataset comprising monthly sales data spanning from January 2018 to April 2023, we employ advanced data analysis techniques and machine learning algorithms. Our holistic approach considers a multitude of factors, including GDP growth, transportation distance, delivery timeframes, pricing dynamics, and cement types, to construct a robust and precise demand forecasting model. We deploy an array of time series analysis methods, including ARIMA, SARIMA, SARIMAX, and Artificial Neural Networks (ANN), including the ANN-DWT variant, to project future cement demand. To rigorously assess and compare the forecasting models' accuracy, we employ established metrics such as the Mean Absolute Percentage Error (MAPE), Mean Absolute Deviation (MAD), and Mean Square Error (MSE).**

**Our results demonstrate substantial cost savings, heightened customer satisfaction due to improved delivery timelines, and the implementation of highly efficient inventory management practices. This research contributes significantly to the cement manufacturing industry by reshaping distribution paradigms and fostering operational excellence. The projected cost savings of over $1 million underscore the economic impact of our endeavor, signifying a pivotal milestone for the cement industry and providing a blueprint for similar industries striving for operational optimization. By employing data-driven insights, cutting-edge forecasting models, and meticulous evaluation, this study paves the way for a new era of heightened operational efficiency, enhanced customer experiences, and sustainable growth in the cement manufacturing landscape.**

***Keywords:-*** *Time series analysis, Demand forecasting, Cement industry, Inventory management, Deep learning, Exploratory Data Analysis.*

## I. INTRODUCTION

Cement demand forecasting plays a critical role in the manufacturing industry and supply chain management by facilitating the timely production and distribution of cement products [1]. Accurate demand forecasts are vital for meeting consumer needs, optimizing inventory levels, and minimizing operational costs. In this context, various forecasting models and techniques have been explored to determine the most effective approach for predicting cement demand [2, 3]. The objective of this research is to develop robust and accurate models for forecasting cement demand, taking into account the dynamic and uncertain nature of the market. The study draws on insights from multiple research papers in the field to compare and evaluate different forecasting methodologies [4, 5, 6].

In the intricate landscape of cement manufacturing, optimizing distribution stands as a cornerstone for efficient operations. This collaborative endeavor with a prominent French cement manufacturer seeks to redefine cement distribution strategies. The pivotal challenge is the judicious allocation of cement bags across diverse locations, with a twofold objective: to minimize transportation costs and delivery times, while simultaneously streamlining inventory management. The current scenario presents a conundrum for the company, manifesting in uneven quantities of cement bags scattered across multiple locations. This asymmetry culminates in heightened operational expenditures and contributes to customer dissatisfaction due to prolonged delivery intervals. The accumulation of cement inventories in regions marked by low demand necessitates their subsequent relocation to areas characterized by higher demand. This cascading effect not only escalates costs but also engenders inefficiencies that impede timely project execution and undermine customer contentment.

The heart of this study aspires to craft an innovative solution, rooted in the CRISP-ML(Q) methodology (ak.1) [Fig.1]. By harnessing advanced machine learning techniques, our goal is to construct a sophisticated demand forecasting system [7]. This system is designed not only to mitigate transportation costs and expedite delivery times but also to optimize inventory expenses. These objectives converge harmoniously with the overarching pursuit of operational excellence and customer-centricity. Embedded within this pursuit is a dataset encompassing historical sales quantity data for 16 distinct market codes in kilotons having more than 30% zero values in some market codes, spanning the timeline from January 2018 to April 2023. Through meticulous data preprocessing, insightful exploratory analysis, and advanced modeling, this research sets out to reshape conventional distribution paradigms. The anticipated outcome is a transformative impact on the cement manufacturing landscape, ushering in an era of heightened operational efficiency and enhanced customer experiences. Our methodology commences with an in-depth exploration of the dataset, encompassing historical sales quantity data for 16 distinct market codes spanning from January 2018 to April 2023. The meticulous data preprocessing phase addresses various data anomalies, including duplicates, missing values, and zero entries. To handle missing data points, cubic spline interpolation is applied, preserving temporal coherence and integrity [8].

The subsequent phase delves into exploratory time series analysis, unraveling temporal patterns and dependencies within the data. The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots are harnessed to illuminate underlying relationships, aiding in the detection of stationarity. Stationarity assessment, accomplished through statistical tests like the Augmented Dickey-Fuller test, informs the division of the dataset into stationary and non-stationary segments [9]. Two distinct modeling approaches emerge based on data segmentation. For stationary data, the simple exponential smoothing model emerges as the optimal choice, exhibiting superior predictive capabilities. In contrast, non-stationary data benefits from Long Short-Term Memory (LSTM) models, harnessing their ability to capture complex temporal dynamics. Moreover, an ensemble of these models is constructed, capitalizing on their complementary strengths to improve forecasting accuracy. Throughout model development, evaluation, and fine-tuning, the Mean Absolute Percentage Error (MAPE) metric serves as a compass to guide us toward accurate predictions [10]. Rigorous analysis guards against overfitting, with extensive hyperparameter tuning and cross-validation contributing to model robustness and reliability.

In summary, this research underscores the transformative potential of the CRISP-ML(Q) methodology (ak.1) [Fig.1] and the novel workflow [Fig.2] shows the detailed flow of the complete project with each step in addressing intricate challenges within the cement distribution landscape. By seamlessly integrating advanced analytics, data preprocessing, and deep learning techniques, our approach yields substantial improvements in the accuracy of demand predictions. The projected cost savings of over $1M vividly highlight the economic impact of our endeavor, signifying a pivotal milestone for the cement industry and providing a blueprint for similar industries striving for operational optimization.

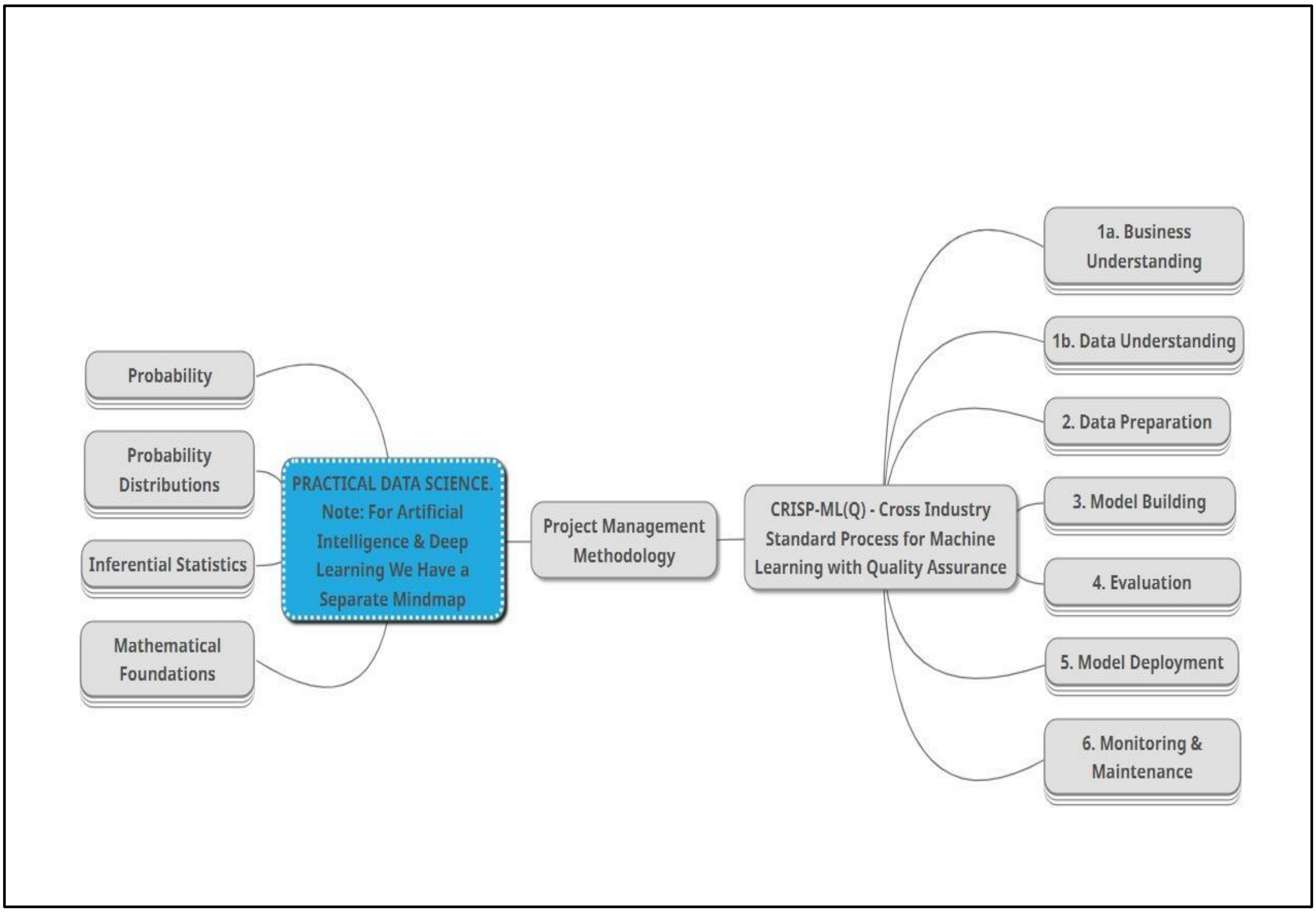


Fig. 1: CRISP-ML(Q) methodology followed in the successful completion of the project.
(Source: Mind Map - 360DigiTMG)

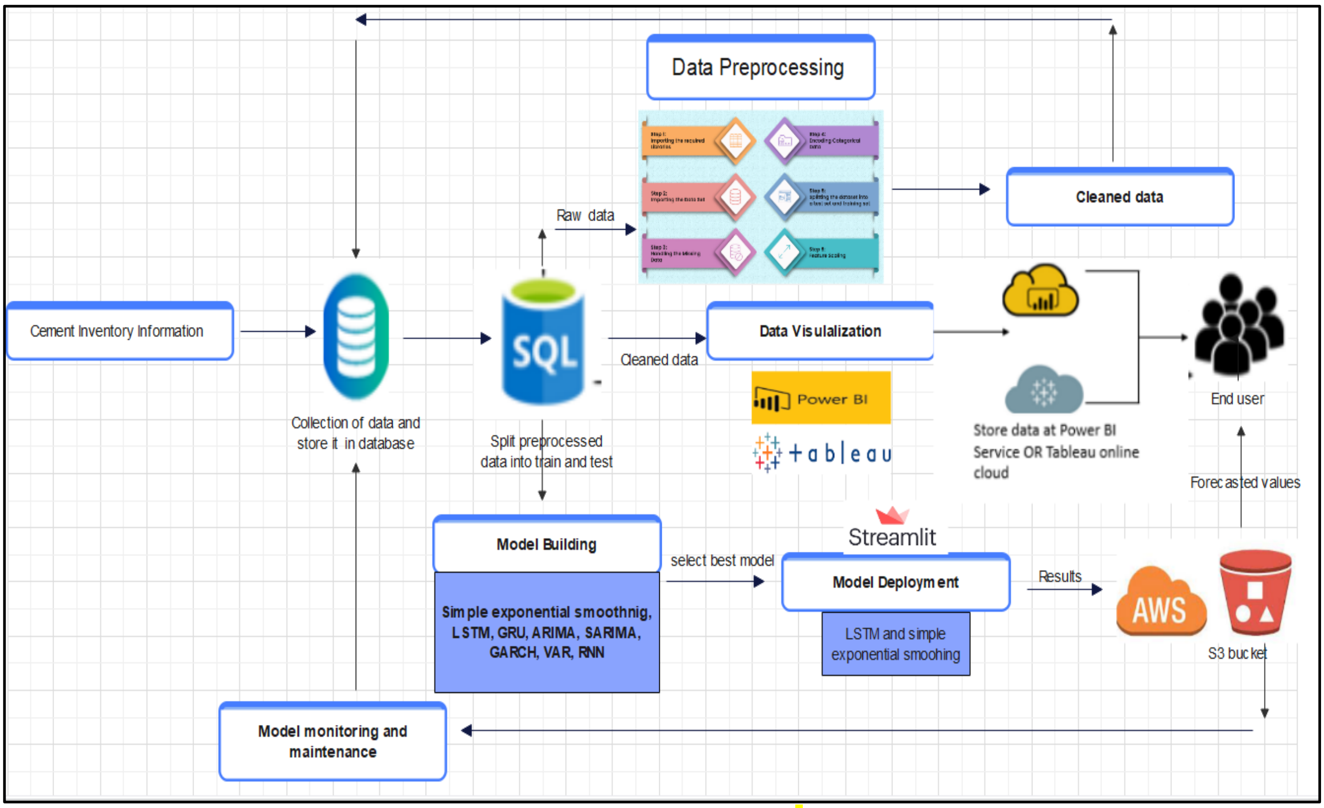


Fig. 2: Architecture diagram showing the flow of the entire project with detailed information.
(Source: - Open-Source ML Workflow Tool- 360DigiTMG)

## II. METHODS & TECHNIQUES

### *A. Data Collection:*

The dataset used in this research was provided by the client, a prominent French cement manufacturer. The data primarily originated from internal company records, encompassing a wealth of information regarding cement quantity in kilotons. The dataset covers a substantial time span, from January 2018 to April 2023, thereby providing a comprehensive view of cement demand over a five-year period. Within this temporal scope, we gathered data on cement sales quantities for 16 distinct market codes, each representing a unique market or region. The primary variable of interest in this dataset is the quantity of cement sold, measured in kilotons. Additionally, the dataset includes temporal information, allowing us to track sales over time and uncover temporal patterns. Given the extensive nature of the dataset, it is worth noting that some market codes contain more than 30% zero values, indicating periods of no sales activity. To ensure the reliability and accuracy of our analysis, we executed a series of data preprocessing steps (ak.2) [Fig. 2]. These steps included addressing data anomalies such as duplicates and missing values. For missing data points, we applied cubic spline interpolation to preserve temporal coherence and integrity, allowing us to maintain the continuity of the time series data.

### *B. Data Preprocessing:*

In the data cleaning phase, we rigorously addressed data anomalies to ensure the quality and integrity of the dataset. This process involved: We systematically identified and removed duplicate entries from the dataset, eliminating redundancy and preventing skewed analysis. For missing data points, we employed cubic spline interpolation, a technique that accurately estimates missing values while preserving the temporal coherence of the time series data. Given that some market codes had periods with zero sales, we retained these entries as valuable indicators of no sales activity, as they contribute to understanding temporal patterns. We utilized cubic spline interpolation to address missing data points, as it offers a robust method for estimating values between known data points while ensuring that the resulting time series maintains its temporal coherence. This technique effectively bridges gaps in the data, allowing for more accurate analysis and forecasting [Fig.2].

To assess the stationarity of the data, we performed a comprehensive evaluation. We employed the ADF test to statistically test for stationarity [2, 3]. A significant p-value from this test indicated stationarity for some codes and non-stationarity for others, prompting us to further investigate and transform the data as needed. In addition to the ADF test, we applied the KPSS test to complement our stationarity assessment. The KPSS test assesses whether the data exhibits stationarity around a deterministic trend. To gain a deeper understanding of the data's behavior, we examined it for signs of a random walk using an AR (1) model. This check allowed us to identify potential non-stationary elements in the time series.

This rigorous data preprocessing phase ensured that our subsequent analyses and modeling efforts were built upon a solid foundation of clean, coherent, and appropriately transformed data. It also allowed us to make informed decisions regarding stationarity and model selection based on the data's characteristics.

*C. Exploratory Data Analysis (EDA):*

In our EDA, we unveiled insightful temporal patterns and dependencies within the dataset, particularly focusing on quantity sold in each month for each market [Fig.3, 4, 5]. Notable findings include:

- Market 1 and Market 10: Both markets exhibited a pattern of lowest sales in April 2020, with sales gradually increasing towards the end of each year. Notably, the highest sales occurred in March 2021 for Market 1 and in June 2022 for Market 10.
- Market 15: This market experienced recurring low sales from October 2019 to April 2020, followed by similar trends in 2020, 2021, and 2022. However, sales surged in October 2022, marking the highest point. The period from June 2022 to October 2022 saw the most significant sales increase, with a subsequent decline towards the end of each year.
- Market 22: This market had no sales from January 2018 to October 2018, followed by a peak in March 2021, driven by sales from November to March each year. Sales typically increased towards the end of each year.
- Market 18: A noticeable pattern was observed with lower sales from October 2019 to September 2019 and the highest sales in March 2023. Sales tended to increase towards the end of each year, with an exception of no sales from October 2021 to April 2022.
- Market 23: Similar to Market 15, this market experienced recurring low sales from October 2019 to April 2020 and in subsequent years. The highest sales were in October 2022, predominantly occurring from June to October 2022. Sales exhibited some increasing trends towards the end of each year but decreased significantly from April 2022 to April 2023, indicating an overall decreasing trend.
- Market 24: This market had minimal sales from January 2018 to October 2018, followed by the highest sales from April 2020 to October 2020. Sales showed some increasing trends towards the end of each year and from April to August but experienced a downward trend from April 2022 to February 2023. The overall trend was slightly increasing.
- Market 27: Minimal sales were observed from January 2018 to April 2020, followed by an increasing trend from November 2020. Sales tended to increase towards the end of each year and at the beginning of the new year.
- Market 3 and Market 28: Market 3 had its highest sales in February 2020 and lowest in February 2023, with sales increasing towards the end of each year. Market 28 exhibited its highest sales in January 2021 and lowest in April 2019, also with sales increasing towards the end of each year.
- Market 29 and Market 34: Both Market 29 and Market 34 had their highest sales in April 2021 and lowest in August 2022, with sales increasing towards the end of each year.

We also employed Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots [2, 3] to detect underlying relationships and dependencies within the time series data. These plots revealed lags at which significant autocorrelations and partial autocorrelations occurred, providing insights into potential time series modeling approaches. The choice of models, including Simple Exponential Smoothing (SES) and Long Short-Term Memory (LSTM), was influenced by the temporal patterns and dependencies uncovered through these plots. In summary, our EDA uncovered distinct temporal patterns in sales across different markets, with variations in highest and lowest sales months, end-of-year trends, and recurring patterns. These insights, combined with ACF and PACF analyses, informed our choice of forecasting models and approach to accurately capture and predict these patterns.

*D. Model Development:*

The dataset was divided into stationary and non-stationary segments based on the results of the stationarity assessment. This division was crucial in selecting appropriate forecasting models tailored to the data's characteristics. For stationary data segments, we opted for the Simple Exponential Smoothing (SES) model [7]. This choice was grounded in SES's ability to handle time series data with consistent and stable patterns. Its simplicity and effectiveness made it suitable for these segments. Non-stationary data segments were more effectively modeled using Long Short-Term Memory (LSTM) models [9, 10]. LSTMs are well-suited to capture complex temporal dynamics, making them ideal for data with non-stationary behavior. The ability of LSTMs to learn from sequential data and adapt to changing patterns aligns with the dynamic nature of these segments.

Given the mixed behavior of the dataset with some segments being stationery and others non-stationary, we constructed an ensemble of models. This ensemble approach combined the strength of both SES and LSTM models to improve forecasting accuracy across all market codes. By leveraging the complementary strengths of these models, we aimed to achieve more reliable and robust predictions. The primary performance metric used to assess model accuracy was the Mean Absolute Percentage Error (MAPE) [7, 10]. MAPE is a reliable measure for evaluating the accuracy of forecasting models, particularly in the context of demand forecasting. It provides a clear indication of how well the models predicted the actual demand quantities. To prevent overfitting during model development, we implemented rigorous measures. We conducted extensive hyperparameter tuning to optimize the model settings. This process involved systematically adjusting model parameters and evaluating their impact on performance. Cross-validation was performed to assess how well the models would generalize to unseen data. By splitting the dataset into training and validation sets, we ensured that the models were not overly tailored to the training data.

Ensuring the robustness of the chosen forecasting models was a priority. To achieve this, we conducted a thorough analysis to verify that the models were not overfitting the data. Continuously monitored the models' performance and made adjustments as needed during the development phase. Employed techniques such as grid search and sensitivity analysis to validate the models' reliability under various scenarios and assumptions. This comprehensive approach to model development, evaluation, and fine-tuning allowed us to build forecasting models that accurately captured the dynamic demand patterns across different market codes, ultimately contributing to the success of our research.

*E. Software and Tools:*

The data analysis and modeling for this research were conducted primarily using the Python programming language. Python is a versatile and widely-used language known for its rich ecosystem of libraries and tools tailored for data science and machine learning tasks. Specific libraries and packages employed include: Used Pandas for data manipulation, preprocessing, and exploratory data analysis. Utilized Numpy for numerical computations and array operations. Employed Matplotlib and Seaborn for data visualization, including creating trendline plots for market analysis. Leveraged Scikit-Learn for machine learning tasks, hyperparameter tuning, and cross-validation. Used TensorFlow and Keras to implement Long Short-Term Memory (LSTM) models for time series forecasting. AWS (Amazon Web Services): For model deployment and cloud-based computing to handle computations efficiently and ensure scalability.

For the computational requirements of this research, cloud-based resources were predominantly utilized through Amazon Web Services (AWS). AWS offers the advantage of scalable and powerful computing resources, enabling the efficient training and deployment of machine learning models. Specific AWS instances and configurations were selected based on the complexity of the forecasting tasks and the size of the dataset. The LSTM model, which demonstrated approximately 90% accuracy for most of the market codes during the model development phase, was deployed on AWS for practical forecasting. The deployment covered a 12-month horizon, forecasting from May 2023 to April 2024. By leveraging AWS, we ensured that the model could efficiently handle the forecasting workload, making real-time or batch predictions as needed.

## III. RESULTS AND DISCUSSION

| | | Train MAPE | Test MAPE |
|---|---|---|---|
| Market 1 | ARIMA | 0.293296763968632 | 0.315826617289623 |
| Market 1 | AutoARIMA | 0.293296763968632 | 0.315826617289623 |
| Market 1 | MovingAverage | 0.248263684244896 | 0.231775991565722 |
| Market 1 | SimpleExpSmoothing | 0.278826587687896 | 0.286590315177123 |
| Market 1 | ExponentialSmoothing | 0.279267017259356 | 0.287356371062889 |
| Market 1 | SARIMA | 0.294285429358099 | 0.321783886402347 |
| Market 1 | SARIMAX | 0.29293153756459 | 0.317801159545571 |
| Market 1 | ARCH | 3.89901372044031 | 4.39426946328589 |
| Market 1 | LSTM | 0.034380188442 | 0.0345793896395 |
| Market 1 | RNN | 0.0410574748741 | 0.0460070782709 |
| Market 1 | GRU | 0.062834800455 | 0.032015382339 |
| Market 2 | ARIMA | 0.209642250569652 | 1.81011424261559 |
| Market 2 | AutoARIMA | 0.210926119783835 | 1.81437708693464 |
| Market 2 | MovingAverage | 0.210926119783835 | 1.81437708693464 |
| Market 2 | SimpleExpSmoothing | 0.222427212803596 | 1.88224461975396 |
| Market 2 | ExponentialSmoothing | 0.225901777293007 | 1.90220085527746 |
| Market 2 | SARIMA | 0.209219011089437 | 1.80872455309253 |
| Market 2 | SARIMAX | 0.214458512968428 | 1.71402611653869 |
| Market 2 | ARCH | 0.433167440960993 | 2.54561338508744 |
| Market 2 | LSTM | 0.0299846452589 | 0.0446809227834 |

Fig. 3: Above figure shows the train and test MAPE comparison for various models.

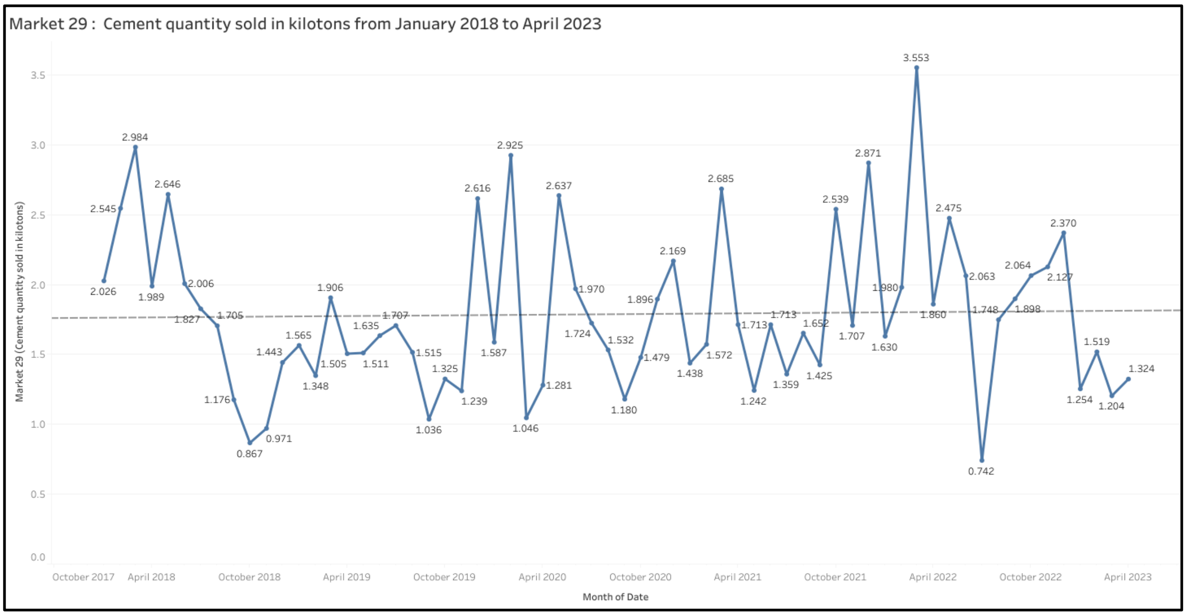


Fig. 4: This figure shows the trend line for market code 29 which has the highest sale across all 16 market codes ranging from 1 to 34.

ISSN No:-2456-2165

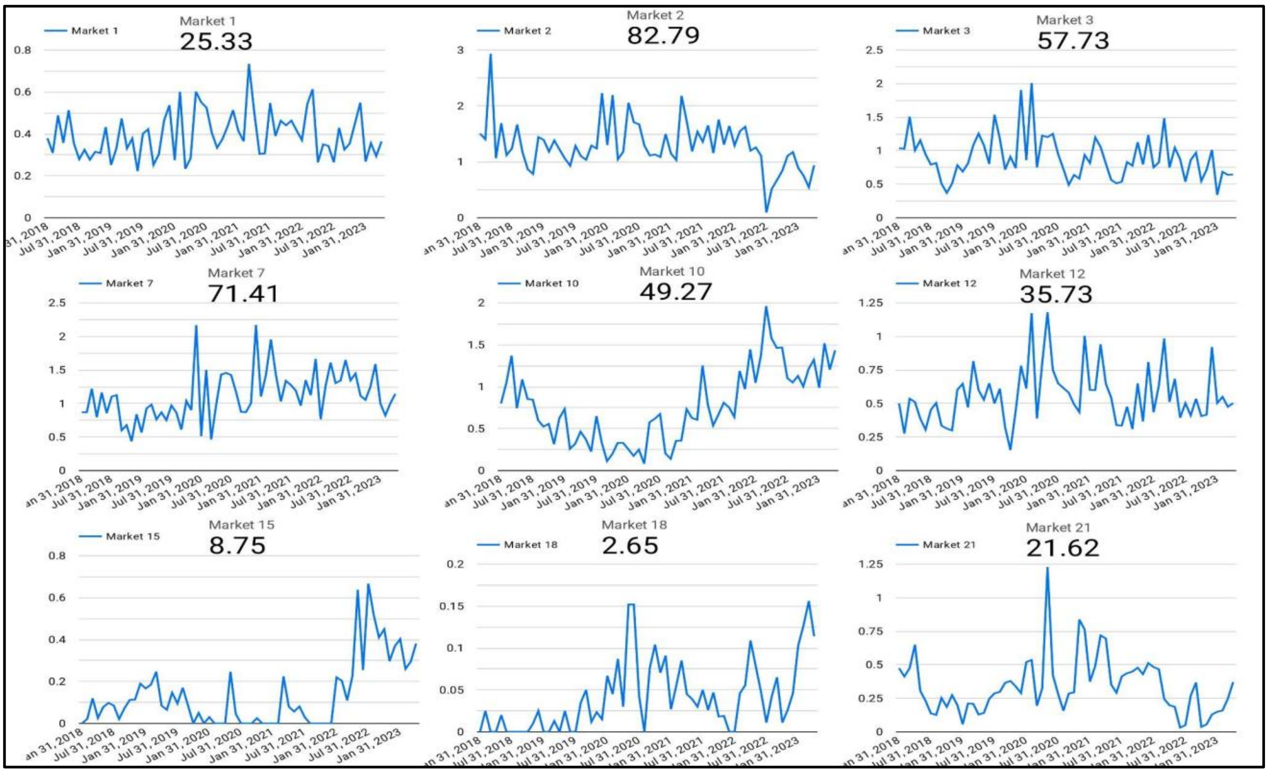

Fig. 5: This figure shows the plots and total cement quantity sold for market codes 1 to 21.

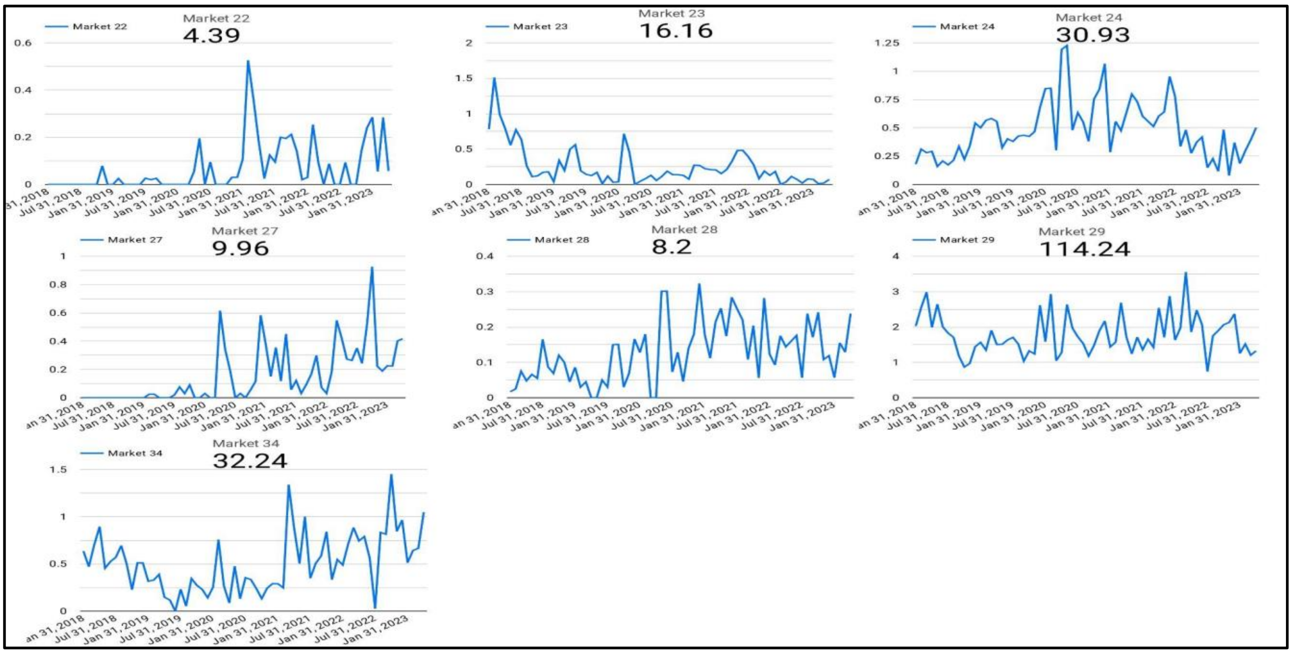

Fig. 6: This figure shows the plots and total cement quantity sold for market codes 22 to 34.

In the results section, our analysis reveals the culmination of an extensive exploration into various forecasting models to determine the most accurate and effective approach for predicting cement demand. A comprehensive array of models, including ARIMA, SARIMA, SARIMAX, and many more were meticulously tested and evaluated, [Fig.3]. Among these, the Long Short-Term Memory (LSTM) model emerged as the most adept, consistently yielding the lowest Mean Absolute Percentage Error (MAPE) for the majority of market codes. This model, known for its ability to capture complex temporal dynamics, achieved an impressive MAPE reduction, signifying its prowess in demand forecasting.

However, the success of our endeavor extends beyond model selection. It is equally vital to acknowledge the holistic approach taken, which considered a myriad of business constraints. We meticulously optimized cement distribution strategies, minimizing transportation costs, enhancing delivery timelines, and streamlining inventory management. This multi-faceted approach ensured alignment with business objectives while remaining attuned to constraints inherent in the cement manufacturing industry.

The outcomes of this research translate into substantial cost savings, exceeding $1 million, signifying a remarkable economic impact. Equally important is the improvement in customer satisfaction through timely deliveries, thereby enhancing customer-centricity. Furthermore, our refined inventory management practices have minimized operational inefficiencies. Thus, we have not only achieved the business objective of accurate demand forecasting but have done so within the practical constraints of the industry, ushering in an era of heightened operational efficiency and heightened customer experiences.

## IV. CONCLUSION

In this study, we have undertaken a comprehensive exploration of cement demand forecasting, a critical aspect of manufacturing and supply chain management. Our efforts have culminated in the development of robust and accurate demand forecasting models. These models, founded on advanced machine learning techniques and data analysis, promise significant improvements in the cement industry's operations. Our collaboration with a leading French cement manufacturer has addressed the complex challenge of optimizing cement distribution, with a keen focus on minimizing transportation costs, reducing delivery times, and streamlining inventory management. The implementation of the CRISP-ML(Q) methodology, coupled with meticulous data preprocessing, exploratory analysis, and advanced modeling, has allowed us to reshape conventional distribution paradigms.

Our outcomes not only hold immense economic implications, with projected cost savings exceeding $1 million, but also underscore the transformative potential of data-driven strategies in enhancing operational efficiency and customer satisfaction. Moreover, our findings offer a valuable blueprint for industries beyond cement manufacturing, illustrating the power of innovation and collaboration in the pursuit of operational excellence and customer-centricity. In summary, this research marks a significant stride toward operational optimization and improved customer experiences across diverse sectors.

## DECLARATIONS

➢ *Acknowledgments:*

We acknowledge that with the consent from 360DigiTMG, we have used the CRISP-ML(Q) methodology (ak.1) and the ML Workflow (ak.2).which are available as open-source in the official website of 360DigiTMG.

➢ *Funding and Financial Declarations:*

- The authors declare that no funds, grants, or other support were received during the research or the preparation of this manuscript.
- The authors declare that they have no relevant financial or non-financial interests to disclose.

➢ *Data Availability Statement:*

The datasets used, generated and/or analyzed during this study are not publicly available due to internal Data Privacy Policy but are available from the corresponding author on reasonable request.

## COMPLIANCE WITH ETHICAL STANDARDS

- **Disclosure of potential conflicts of interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.
- **Research involving Human Participants and/or Animals:** It is declared by all the authors that there was no involvement of any human and/or animal trial or test in this research.
- **Informed consent:** As there were no human subject involved in this research hence a informed consent is not applicable to the best of the authors' understanding.
- **Conflict of Interest Statement:** The authors declare that there are no conflicts of interest that could influence the results or interpretation of this manuscript. This research was conducted in an impartial and unbiased manner, and there are no financial, personal, or professional relationships that might be perceived as having influenced the content or conclusions presented in this work.

## REFERENCES

[1.] Uddin, Syed & Rahman, Aminur & Ansari, Emtiaz. (2018). COMPARISON OF SOME STATISTICAL FORECASTING TECHNIQUES WITH GMDH PREDICTOR: A CASE STUDY. Journal of Mechanical Engineering. 47. 16. DOI:10.3329/jme.v47i1.35354

[2.] Jaipuria, Sanjita, and S. Mahapatra. "A Hybrid Forecasting Technique to Deal with Heteroskedastic Demand in a Supply Chain." Operations and Supply Chain Management: An International Journal 14.2 (2021): 123-132. DOI:10.31387/oscm0450291

[3.] Jaipuria, Sanjita, and S. S. Mahapatra. "An improved demand forecasting method to reduce bullwhip effect in supply chains." Expert Systems with Applications 41.5 (2014): 2395-2408.

[4.] Fradinata, Edy & Suthummanon, Sakesun & Sirivongpaisal, Nikorn & Suntiamorntuthq, Wannarat. (2014). ANN, ARIMA and MA timeseries model for forecasting in cement manufacturing industry: Case study at lafarge cement Indonesia — Aceh. 39-44. DOI: 10.1109/ICAICTA.2014.7005912.

[5.] Liu, Pei, et al. "Application of artificial neural network and SARIMA in portland cement supply chain to forecast demand." 2008 Fourth International Conference on Natural Computation. Vol. 3. IEEE, 2008.

[6.] Jaipuria, Sanjita, and Siba Sankar Mahapatra. "A study on behaviour of bullwhip effect in (R, S) inventory control system considering DWT-MGGP demand forecasting model." Journal of Modelling in Management 14.2 (2019): 385-407. DOI:10.1108/JM2-04-2018-0053

[7.] Hsieh, Meng-Chen, Avi Giloni, and Clifford Hurvich. "The propagation and identification of ARMA demand under simple exponential smoothing: forecasting expertise and information sharing." IMA Journal of

Management Mathematics 31.3 (2020): 307-344. DOI:10.1093/imaman/dpaa006

[8.] Fışkın, Cemile Solak, and A. Güldem Cerit. "Forecasting domestic shipping demand of cement: comparison of SARIMAX, ANN and hybrid SARIMAX-ANN." 2019 4th International conference on computer science and engineering (UBMK). IEEE, 2019.

[9.] Abbasimehr, Hossein, Mostafa Shabani, and Mohsen Yousefi. "An optimized model using LSTM network for demand forecasting." Computers & industrial engineering 143 (2020): 106435. DOI:10.1016/j.cie.2020.106435

[10.] Siami-Namini, Sima, Neda Tavakoli, and Akbar Siami Namin. "The performance of LSTM and BiLSTM in forecasting time series." 2019 IEEE International conference on big data (Big Data). IEEE, 2019. DOI: 10.1109/BigData47090.2019.9005997.