**Sales Volume Prediction in Curd Industry: A Data Science Approach**

**1. Introduction**

**1.1 Problem Statement and Background**

Mother Dairy’s curd operates in a complex environment where accurate prediction of daily sales volume is crucial for operational success. The perishable nature of the product, combined with variable demand patterns influenced by multiple external factors, creates a significant challenge for inventory management and production planning. Traditional forecasting methods have proven inadequate in capturing the intricate relationships between various factors affecting sales volume, leading to substantial operational inefficiencies and potential revenue loss. This project was initiated to address these challenges through the application of advanced data science techniques and methodologies.

**1.2 Project Objectives**

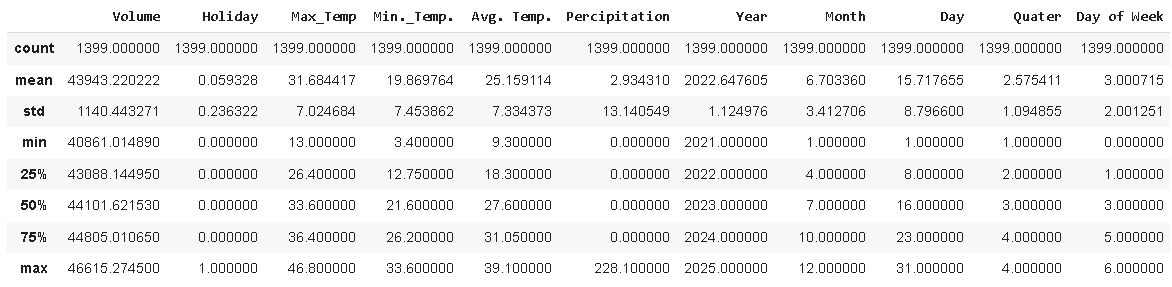
* Develop accurate volume prediction models.
* Analyse the impact of external factors on sales volume.
* Compare different modelling approaches.
* Provide implementable solutions for business use.

**2. Data Understanding**

**2.1 Dataset Description**

The dataset contains daily records with the following features:

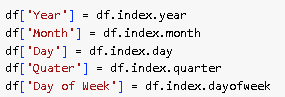
* Date (index): Dates from “01-04-2021” to “28-01-2025”
* Volume (target variable): Sales Volume
* Holiday (binary indicator): 0 = non-Holiday and 1 = Holiday
* Temperature metrics: Maximum Temperature, Minimum Temperature, Average Temperature
* Precipitation: Measure of rain
* Derived temporal features: Year, Month, Day, Quarter, Day of Week (these features were driven from the date column)



**3. Data Preprocessing**

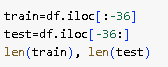
**3.1 Feature Engineering**

* Created temporal features from date index



**3.2 Data Splitting**

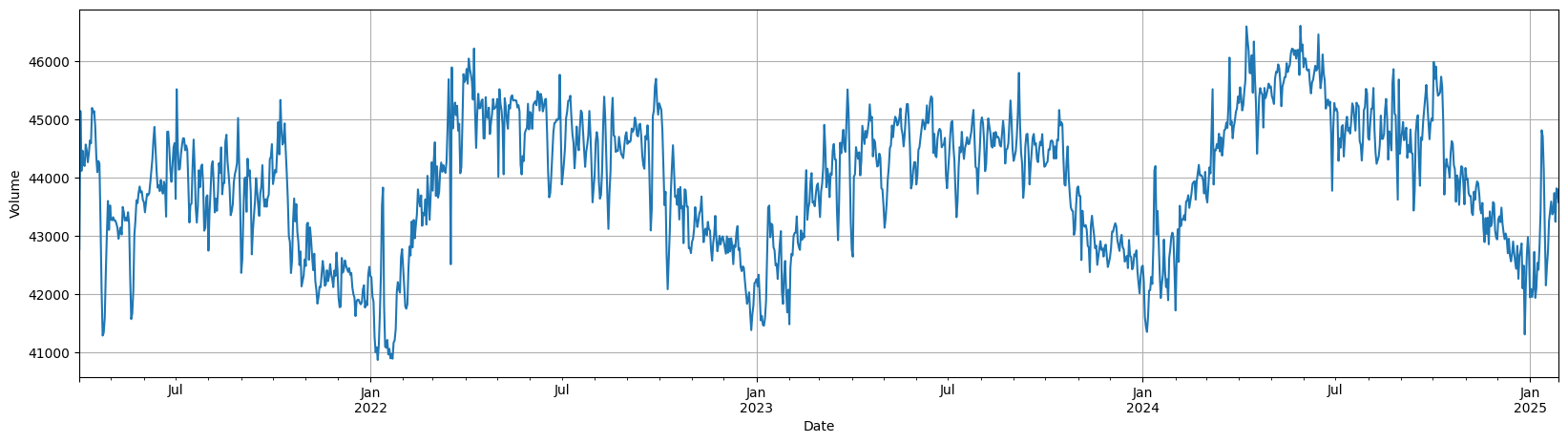
* Training set: 80% of data
* Testing set: 20% of data



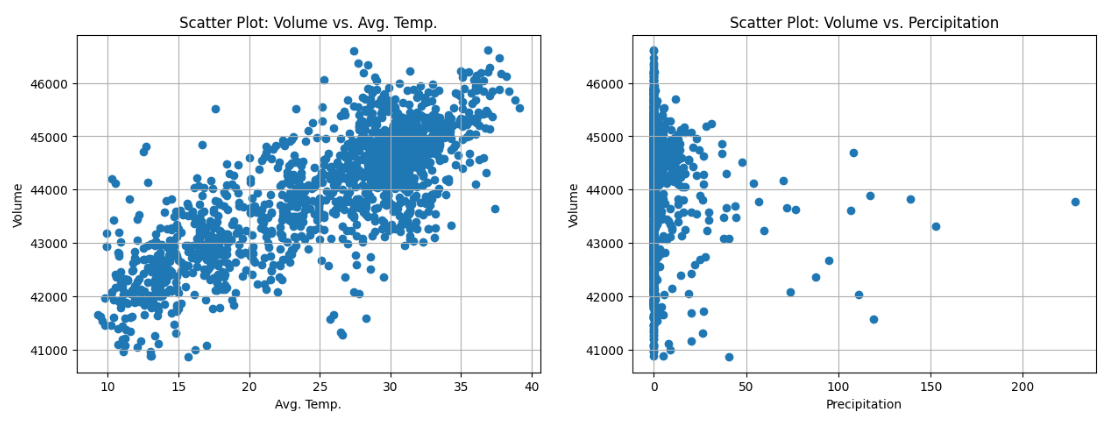
**4. Exploratory Data Analysis (EDA)**

**4.1 Feature Analysis**

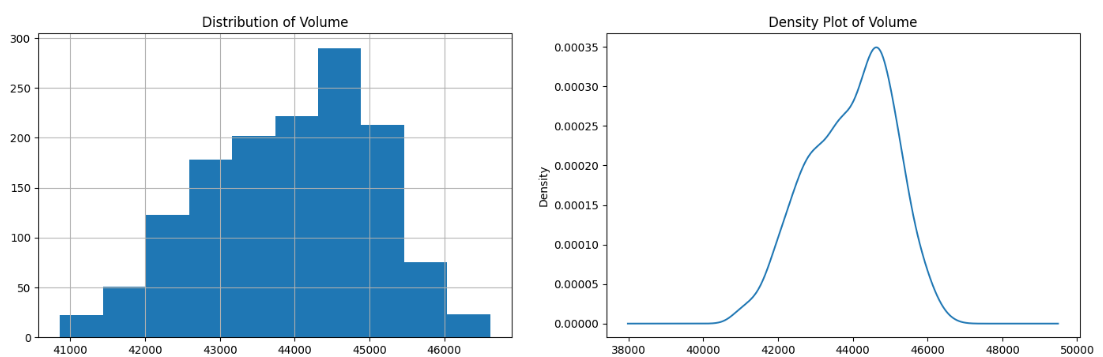
* **Time Series Plot of Volume:** This plot visualizes the trend and seasonality of your target variable ('Volume') over time. It's crucial for understanding the historical patterns in your data, which is vital for forecasting.



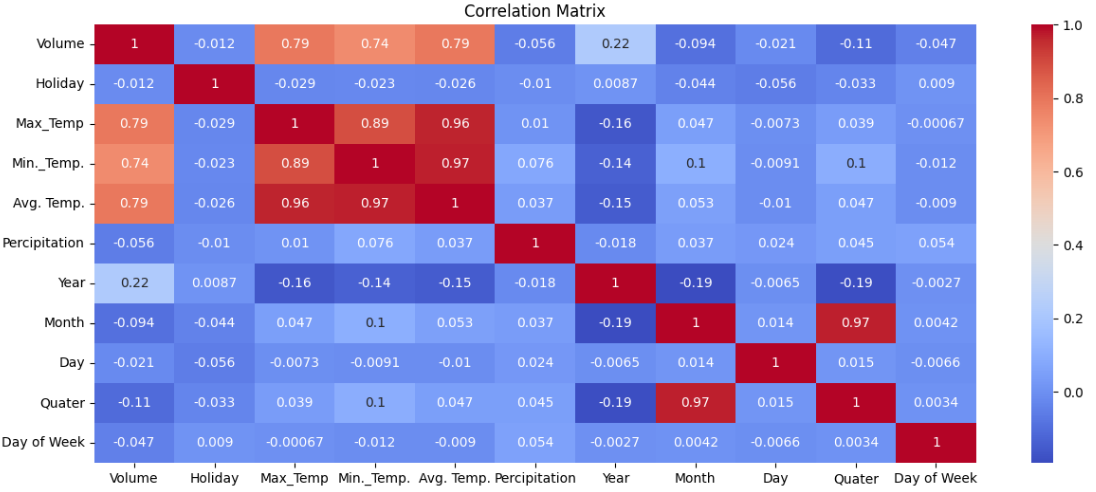
* **Scatter Plots of Volume vs. Key Features:** These plots visualize the relationship between 'Volume' and individual predictor features like 'Avg. Temp.', 'Precipitation'. They help in understanding the nature of the relationships (linear, non-linear) and identifying any potential patterns or clusters.



* **Distribution of Volume:** This shows the frequency or density of different 'Volume' values using histograms or density plots. It helps in identifying outliers, skewness and the overall distribution shape, crucial for choosing appropriate models.



* **Correlation Matrix**: To visualize and understand the linear relationships between numerical features in the dataset, particularly those related to 'Volume'.



\*Darker colours represent stronger correlations.

\*Lighter colours represent weaker correlations.

**5. Augmented Dickey-Fuller (ADF) test**

* This is a statistical test that checks for unit roots in a time series. It's used to determine if a time series is stationary or non-stationary.
* The Augmented Dickey-Fuller (ADF) test is important because it is a crucial tool for determining whether a time series data is stationary or not, which is a key requirement for applying many time series analysis techniques, particularly when forecasting, as most models (SARIMAX, VAR VARMA) rely on stationary data for accurate predictions; identifying non-stationarity through the ADF test allows for necessary data transformations to achieve stationarity before further analysis.





* If a feature fails the Augmented Dickey-Fuller (ADF) test, it means the time series is non-stationary, which can negatively impact forecasting models, especially those that assume stationarity, such as ARIMA or certain deep learning models. To make the series stationary, we should apply differencing, which removes trends and seasonality by subtracting the previous value from the current value: 
* After applying differencing, re-run the ADF test to confirm stationarity before proceeding with modelling.

**6. Model Evaluation**

**6.1 Evaluation Metrics**

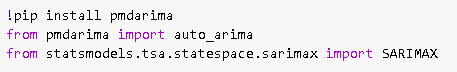
* Root Mean Square Error (RMSE): It is the residuals' standard deviation, or the average difference between the projected and actual values produced by a statistical model.
* Normalized Root Mean Square Error (NRMSE): a metric used to evaluate the accuracy of a predictive model by comparing its predicted values to observed values, but normalized by the range (difference between maximum and minimum) of the observed data.
* Mean Absolute Error (MAE): It is measured as the average absolute difference between the predicted values and the actual values and is used to assess the effectiveness of a model.
* R-squared (R²): It is defined as a number that tells you how well the independent variable(s) in a statistical model explains the variation in the dependent variable.

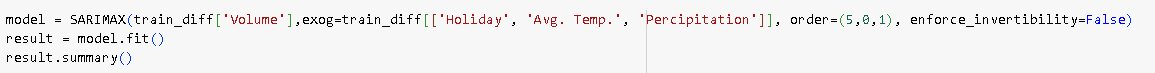
**7. Modelling Approach**

**7.1 Statistical Models**

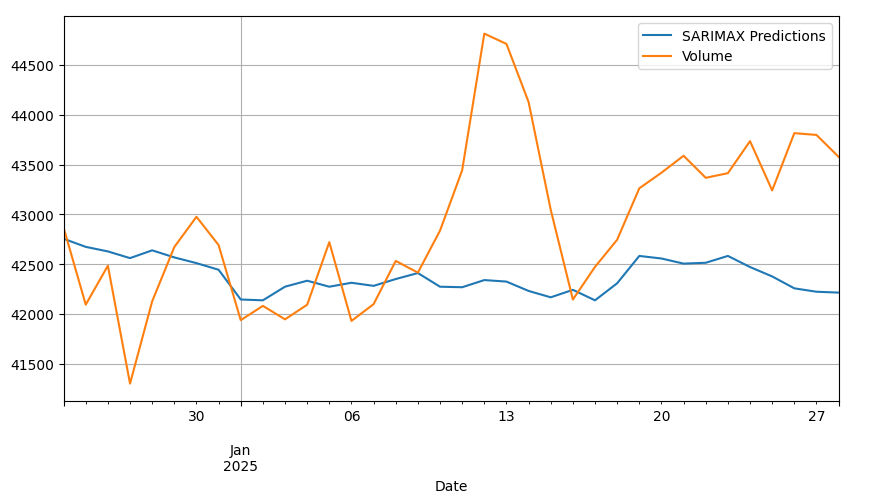
**7.1.1 Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX)**

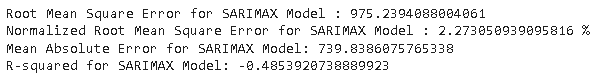
* SARIMAX is an advanced time series forecasting model that extends SARIMA (Seasonal ARIMA) by incorporating exogenous variables (independent predictors). It is widely used for forecasting time series data that exhibit both trend and seasonality while also considering external influences.
* Dependencies:





* Implementation details:
  + Order: (5,0,1)
  + Exogenous variables: Holiday, Temperature, Precipitation
* Hyperparameter optimization through grid search.
* Model diagnostics performed for residual analysis.
* Performance:



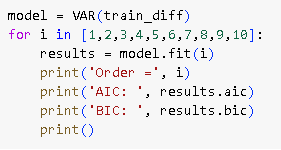


* The relatively high RMSE and NRMSE indicate the presence of some larger prediction errors.

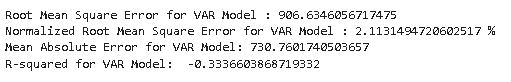
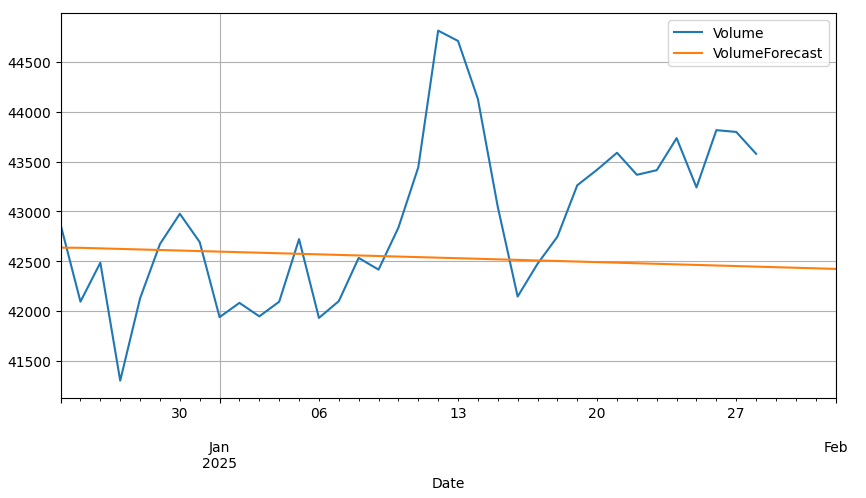
**7.1.2 Vector Autoregressive (VAR)**

* VAR is a statistical model that describes how multiple variables change over time. It's a type of regression model that uses lagged data to predict future outcomes. VAR models are used in many fields, including economics, medicine, and epidemiology.
* Dependencies:





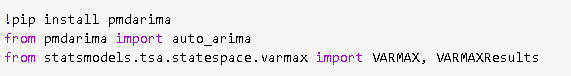
* Lag order selection through information criteria
* Integration with exogenous variables
* Residual analysis for model validation
* Performance:



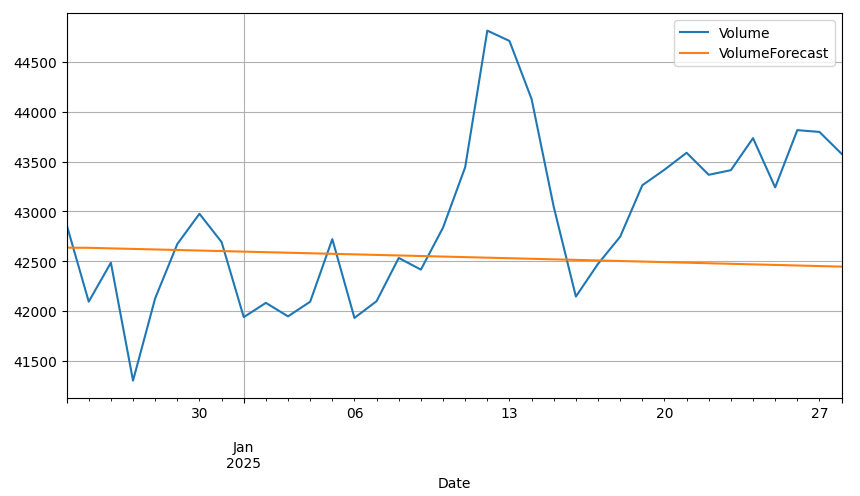
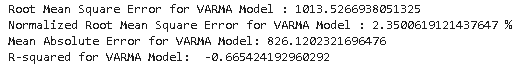
* VAR exhibits poor performance, especially with a negative R-squared value. This model doesn't seem to capture the underlying patterns in the data well.

**7.1.3 Vector Autoregressive Moving Average (VARMA)**

* This is a statistical method used to analyse the dynamic relationships between multiple time series variables, where each variable is explained by its own past values (autoregressive component) as well as a weighted average of past error terms from all the variables in the system (moving average component), allowing for the exploration of how these variables influence each other over time.
* Dependencies:

****

* Performance:

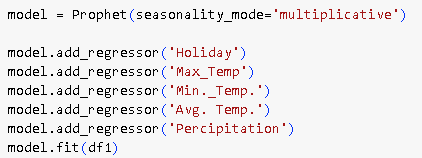
* VARMA's performance is worse than VAR, with higher error metrics and a negative R-squared value. This model also struggles to provide accurate predictions.

**7.2 Facebook Prophet Models**

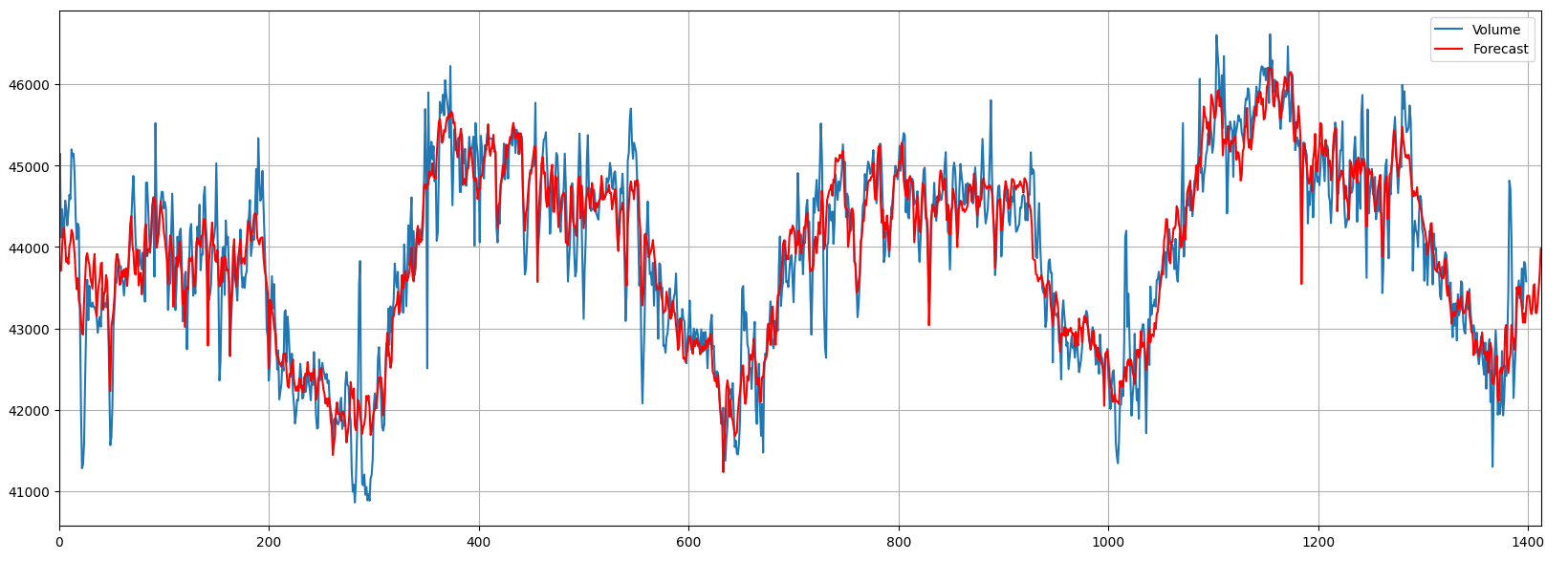
**7.2.1 Prophet Model**

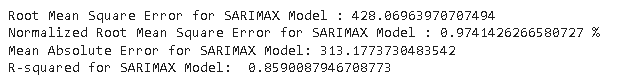
* Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.
* The input to Prophet is always a dataframe with two columns: ds and y and other features or regressors. The ds (datestamp) column should be of a format expected by Pandas, ideally YYYY-MM-DD for a date or YYYY-MM-DD HH:MM:SS for a timestamp. The y column must be numeric, and represents the measurement we wish to forecast.
* Dependencies:





* Focused feature set
* Optimized seasonality settings
* Adding regressors or additional features
* Reduced complexity for better generalization
* Performance:





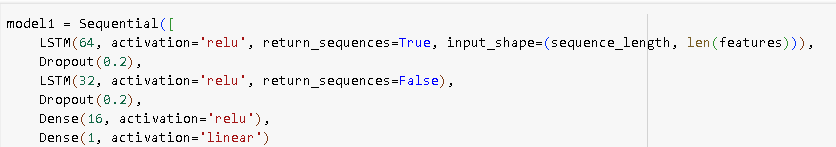
* Prophet shows promising results with a low RMSE and NRMSE, along with a high R-squared. It seems to capture the data patterns effectively.

**7.3 Deep Learning Models**

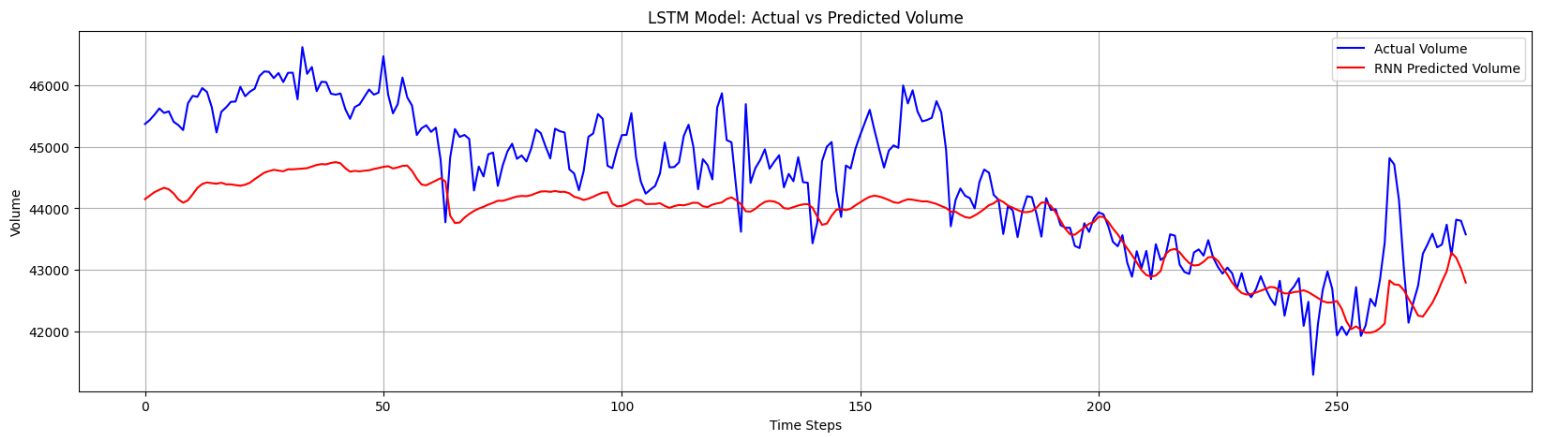
**7.3.1 Long Short-Term Memory (LSTM) based Recurrent Neural Networks (RNN)**

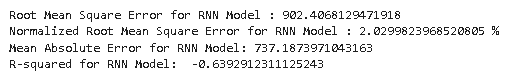
A type of recurrent neural network that utilizes a specialized architecture called "Long Short-Term Memory" to effectively handle long-term dependencies in sequential data, overcoming the limitations of standard RNNs by employing "gates" to control the flow of information and prevent the vanishing gradient problem.

* Dependencies:

* Architecture:
  + Two LSTM layers (64 and 32 units)
  + Dropout layers for regularization
  + Dense layers for final prediction
* Sequence length: 10
* Batch size: 32
* Training epochs: 20
* Performance:

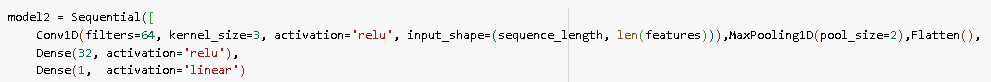




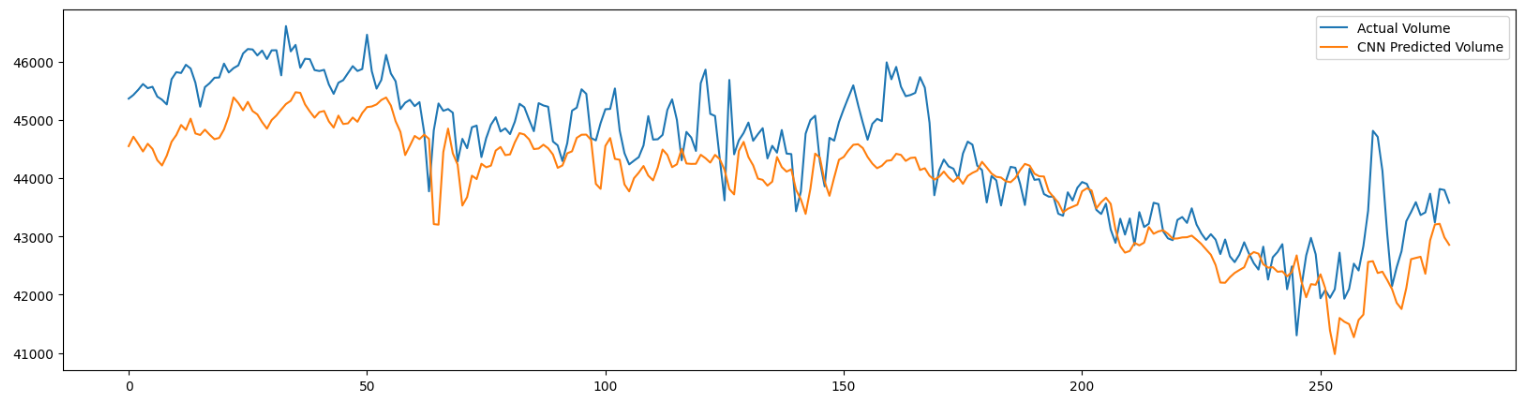
* RNN performs reasonably well, although not as well as the Prophet models. The relatively high RMSE and NRMSE show potential areas for improvement.

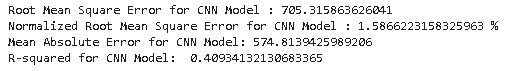
**7.3.2 Convolutional Neural Network (CNN)**

* A CNN (Convolutional Neural Network) model for time series forecasting applies convolutional layers to capture temporal patterns and dependencies in sequential data. Unlike traditional methods like ARIMA or LSTMs, CNNs leverage convolutional filters to extract meaningful features from input time series data, making them effective for forecasting tasks.



* Architecture:
  + 1D Convolutional layers
  + Max Pooling layers
  + Dense layers for prediction
* Filter configuration optimized for temporal patterns
* Training parameters aligned with LSTM
* Performance:

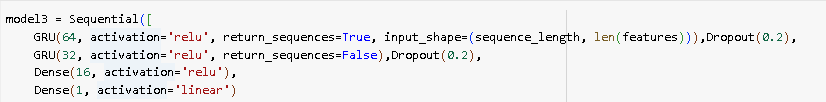




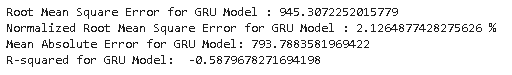
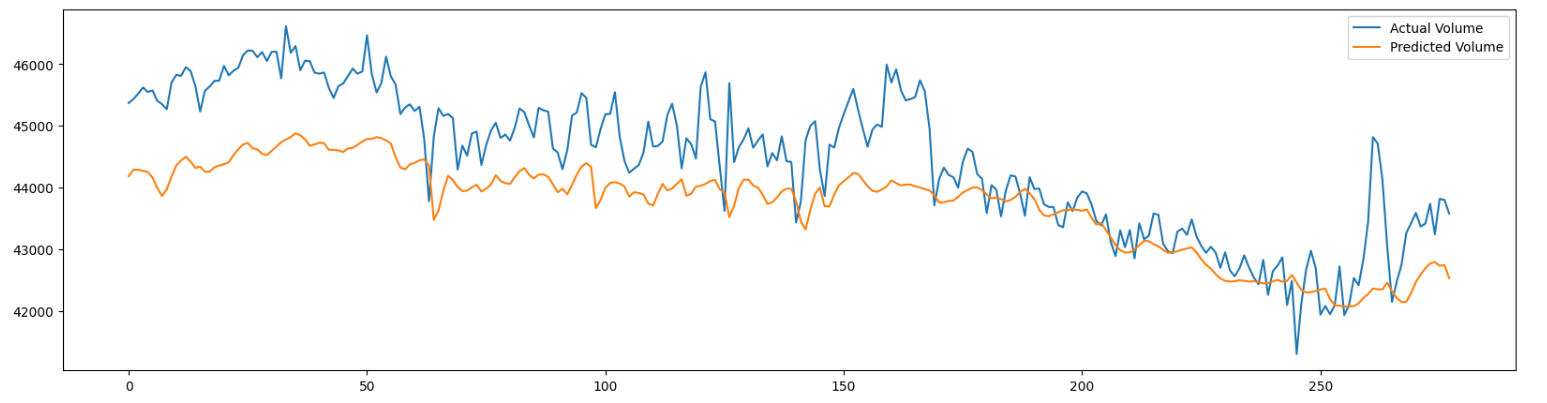
* CNN's performance is slightly worse than RNN, with higher error metrics. It might require further tuning to improve its predictions.

**5.3.3 GRU**

* A Gated Recurrent Unit (GRU) is a type of recurrent neural network (RNN) architecture designed to handle sequential data, such as time series. It is an improvement over traditional RNNs, addressing issues like vanishing gradients and long-term dependencies.

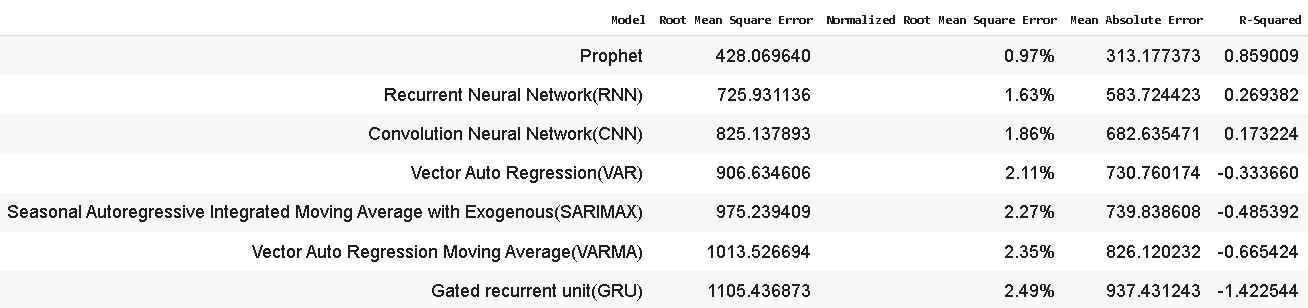
****

* Similar architecture to LSTM
* Modified internal gates for efficiency
* Comparable training parameters
* Performance:



* GRU delivers comparable results to RNN, with similar RMSE and NRMSE values. It demonstrates acceptable performance but has potential for improvement.

**8. Comparing Models & Observation**



* Prophet outperforms all models with the lowest RMSE (428.07) and highest R² (0.859).
* Lowest RMSE (428.07) and MAE (313.18) → More accurate predictions.
* Highest R² (0.859) → Explains 85.9% of the variance, making it the best-fitting model.
* Lowest Normalized RMSE (0.97%), indicating minimal forecasting errors.

The Prophet model demonstrates the best overall performance among the evaluated models, achieving the lowest RMSE, NRMSE percentage and MAE values, while having a highest R-squared value. This suggests that Prophet provides the most accurate predictions with the least error.

**9. Business Insights & Implementation**

The insights derived from this forecasting model can be effectively utilized by curd manufacturers and distributors to optimize their operations, minimize wastage, and improve profitability. Below is key business recommendations based on the analysis:

**1. Demand-Based Production Planning**

* Insight**:** Forecasting allows manufacturers to anticipate spikes in demand, especially during certain seasons, holidays, or temperature changes.
* Actionable Strategy:
  + Adjust production schedules to increase supply during high-demand periods (e.g., summer months, festive seasons) and scale down during low-demand periods to prevent overproduction and wastage.
  + Utilize historical data trends and predicted demand to manage raw material procurement efficiently (e.g., ensuring milk supply aligns with expected curd production).

**2. Inventory & Supply Chain Optimization**

* Insight: Poor demand forecasting can lead to excessive inventory, leading to spoilage and financial losses.
* Actionable Strategy:
  + Implement a just-in-time inventory approach using daily or weekly demand forecasts to prevent overstocking.
  + Work closely with retailers and distributors to align stock levels with expected demand and avoid unnecessary surplus.
  + Use real-time forecasting models to update production levels dynamically based on changes in consumer demand patterns.

**3. Regional Distribution & Expansion**

* Insight: Demand varies by region based on weather, festivals, and consumption habits.
* Actionable Strategy:
  + Identify regions with higher demand and focus distribution efforts accordingly.
  + Use predictive insights to expand into new markets with high curd consumption potential.
  + Optimize logistics routes to ensure faster delivery to high-demand areas, reducing spoilage risks.

**Conclusion**

By leveraging the demand forecasting model, curd manufacturers can shift from reactive to proactive decision-making, reducing waste, optimizing supply chains, and maximizing revenue. Implementing these strategies ensures sustained growth and better market competitiveness.

**9. Final Thoughts**

The project demonstrates a solid understanding of time series forecasting methodologies, making it a valuable contribution to demand prediction in the curd industry. With further refinements and real-world application, it has the potential to drive operational efficiency and revenue optimization.

Report prepared by: Akshat Singh

Date: February 18, 2025