Grocery Demand Prediction

Teja Swaroop Pothala
Student ID:3128337
University of Kansas

Charan Kumar Grandam
Student ID:3129197
University of Kansas

Dinesh Devanaboina
Student ID:3129411
University of Kansas

Venkata Sai Sriram Potluri
Student ID:3126947
University of Kansas

Laxmi Prasanna Kasireddy
Student ID:3146135
University of Kansas

***Abstract*—Predicting a grocery product's demand is a very critical challenge in the retail industry. This explains the performance dynamics of a product and its competitiveness with its rivals in the stores. This project pivots on extracting beneficial insights from a retail sales dataset that covers various products, and supplementing this with additional datasets that will give richer insights that will help a brand or company to plan its production and investing into R&D. With the help of in-depth exploration and effective methods of merging datasets we try to predict the product demand from its historical data. Forecasting Machine Learning Models will be used to predict the demand for grocery products that will take in all the seasonal and non-seasonal time series attributes into account which will participate in the prediction.**

I. Problem Statement and Significance

In today's world with heavy competition in the retail industry with presence of e-commerce and digital marketing it is very challenging for brands to stay top in their domain. Therefore an accurate demand forecasting is required to ensure efficient operations and profitability on the other side of the coin a bad prediction can result in factors like overstocking or understacking, wasted storage, unnecessary markdowns, stockouts etc. Traditional methods struggle to handle the complex interactions of variables that can influence sales of these products.

The increasing availability of data from multiple sources, such as promotional activities, weather conditions, and macroeconomic trends like the country's economical health, presents an opportunity to improve demand prediction accuracy. Trying to merge all these variables to make the prediction more precise and specific. For this problem neural networks are suited because of their ability to model non-linear relationships and learn from large datasets with complex interactions.

II. Related Works

[1] The authors of the paper "A data-driven framework to new product demand prediction: Integrating product differentiation and transfer learning approach" proposed a data-driven framework for predicting demand for new products. This prediction included integration with product differentiation and transfer learning methods. The model they utilized uses a demand differentiation index (DDI) to measure the similarity between existing old products and new products that will get more accurate prediction. Later in the paper the authors introduce a modified version of DDI which is a exponential weighted moving average method. This framework is tested on real-world data from automobile industry which showed promising results.

[2] The authors of the paper "Demand prediction using machine learning methods and stacked generalization" focused mainly on improving demand forecasting by using stacked generalization. This is an ensemble method that can combine prediction from various models. The key algorithms that were introduced are decision trees, random forests and gradient boosted trees. The whole framework integrates the predictions from every algorithm into a meta-learner that results in precise and accurate predictions.

[3] The authors of the paper "Demand prediction, predictive shipping, and product allocation for large-scale e-commerce" basis on optimizing key logistics functions demand forecasting,, predictive shipping and product allocation. The authors proposed a ML framework that enhances these function with the aid of historical sales data, customer behaviour and other factors. So the demand prediction is done by a model that trains on factors like categories, regions, stocking improvisation and shipping efficiency. Factors like allocating products to distribution centers, reduced shipping times and operational costs with the model is used for predictive shipping. And finally the product allocation is achieved by distributing warehouses based on demand forecasts, customer proximity, ensuring quick delivery with lower shipping expenses. All these combinations allows e-commerce platforms to meet customer expectations through data-driven decisions.

[4] The authors of the paper "Demand Prediction and Price Optimization for Semi-Luxury Supermarket Segment" address the challenge of predicting demand and optimizing pricing strategies for high-end supermarket products. The high-end supermarket products are very niche as they have very specific seasonal and highly variable demand that makes it the pricing a crucial factor. The study proposes a decision-support system that integrates machine learning models such as random forests and regression

trees. Also a integer programming model optimizes the pricing based on the forecasts and also balancing demand fluctuations with revenue goals.

III. Dataset Description

There are a total of six datasets in this project. Following is a brief introduction about each of them:

(a) sales.csv
- **id:** An unique identifier for every row in the dataset.
- **unit_sales:** The target column that explains the number of units or weight of the product sold and the demand that we need to predict of how many units might be required or the weight of the product to be kept in stock. Negative value describes the return of the product.
- **onpromotion:** An indication if the product is on discount.
- **item_nbr:** An identifier for the product.
- **store_nbr:** An identifier of the store where the product is sold.
- **date:** Date on the sale of the product.

(b) holidays_events.csv
- **date:** The unique id in the dataset and the date that we are describing if there is a holiday.
- **type:** Description on the kind of holiday and its importance.
- **locale_name:** If the event is a local or at a national level.
- **description:** More detail about the holiday.
- **transferred:** A boolean value if the holiday is moved on the calendar from the day the event officially falls.

(c) items.csv
- **item_nbr:** The unique id for a product.
- **family:** The category of the product.
- **class:** This is a subcategory of the item under the family variable.
- **perishable:** Based on the whether the product has a limited shelf life or not an respective weight is provided to the product.

(d) oil.csv
- **date:** The date of the oil price record.
- **dcoilwtico:** The price of oil (West Texas Intermediate, WTI) on that date, which might be a key economic indicator influencing demand predictions.

(e) stores.csv
- **store_nbr:**Unique identifier for each store.
- **city:** The city where the store is located.
- **state:** The state where the store is located.
- **type:** Store type, which could reflect the size or format for e.g., hypermarket or local.
- **cluster:** Grouping stores into clusters with similar characteristics.

(f) transactions.csv
- **date:** The date of the transaction.
- **store_nbr:** Identifier linking the transaction to a specific store.
- **transactions:** Number of transactions recorded on that date.

IV. Dataset Integration

We extensively analyzed the datasets where we found any outliers in the datasets or if they needed any kind of cleaning before we merged everything into one dataset that could help us train the model to predict the demand:

(a) Sales Dataset
- **Purpose:**The sales dataset (sales.csv) serves as the primary dataset, providing information on the number of units sold for each item in various stores on specific dates.
- **Loading and Initial Inspection:**The dataset contains columns for date, store_nbr, item_nbr, unit_sales, and onpromotion.
- **Key Findings:**
  - **Negative Values:** These represent product returns. It was noted that negative unit_sales needed to be retained as they provide valuable insights into returns.
  - **Missing Data:**The onpromotion column had significant missing values. These were addressed by filling the missing values with False, assuming no promotion when data was missing.
  - **Data Gaps:** No sales data was recorded for December 25th each year, indicating that stores were closed on Christmas. This gap was not imputed, as it reflects a real-world business closure.
- **Preparation:** The date column was converted to a datetime format for easier merging and analysis. This dataset forms the core for subsequent integrations.

(b) Store Dataset
- **Purpose:**The store dataset (stores.csv) provides metadata about the stores where sales occur, such as store numbers and locations.
- **Loading and Initial Inspection:** The dataset includes columns for store_nbr and other store-specific attributes like city, state, and store type.
- **Key Findings:**
  - The store number (store_nbr) acts as the primary key for merging with the sales dataset.
  - The store locations are essential for associating regional and local attributes, such as holidays.
- **Integration:** The store dataset was merged with the sales dataset on the store_nbr column, enriching the sales data with store metadata like city and state.

(c) Item Dataset
- **Purpose:**The item dataset (items.csv) contains metadata about the items sold, such as item categories or families.

- **Loading and Initial Inspection:** The dataset includes item_nbr as the primary key and additional attributes like family and class that categorize the items.
- **Key Findings:**
  - This dataset helps segment sales data by item type or category, enabling category-level analysis.
- **Integration:** The item dataset was merged with the sales dataset using the item_nbr column. This added contextual information about each item, such as its family or class.

(d) Transactions Dataset

- **Purpose:**The transactions dataset (transactions.csv) records the number of transactions that occurred in each store on specific dates. This provides insights into overall store activity.
- **Loading and Initial Inspection:** The dataset contains date, store_nbr, and transactions.
- **Key Findings:**
  - **Trends:**The number of transactions peaked around Christmas Eve. The lowest number of transactions occurred around New Year's Day.
  - **No Missing Values:**The dataset did not have missing values and required minimal cleaning.
- **Integration:** The transactions dataset was merged with the sales dataset on date and store_nbr. This integration helps link the number of transactions to the sales volume, enabling analysis of store traffic alongside sales.

(e) Oil Dataset

- **Purpose:**The oil dataset (oil.csv) tracks daily oil prices, which can influence economic conditions and consumer behavior.
- **Loading and Initial Inspection:** The dataset contains date and dcoilwtico (daily oil price).
- **Key Findings:**
  - **Missing Values:**The dcoilwtico column had missing values. These were filled with the average oil price for the corresponding month.
  - **Economic Insight:**Oil prices can impact inflation, travel, and overall spending behavior.
- **Data Transformation:**Oil prices were categorized into slabs (bins) to facilitate easier analysis of oil price ranges.
- **Integration:**The oil dataset was merged with the sales dataset on the date column. This integration helps analyze the impact of oil prices on sales trends.

(f) Holidays Dataset

- **Purpose:**The holidays dataset (holidays_events.csv) provides information about holidays celebrated at national, regional, and local levels.
- **Loading and Initial Inspection:** The dataset contains date, type, locale, locale_name, description, and transferred.
- **Key Findings:**
  - **Holiday Types:**The dataset includes national, regional, and local holidays.
  - **Mapping Holidays to Stores:**Since the sales data uses store numbers and the holidays data uses cities, a mapping was created from stores to cities using the store dataset. Regional and national holidays were expanded to match relevant cities.
- **Data Transformation:**A new dataset was created to assign holidays to specific cities based on their regions or states.
- **Integration:**The holidays dataset was merged with the sales dataset on date and city. This flagged whether each sales record occurred on a holiday.

(g) Summary

- **Sales Dataset:** The core dataset containing sales information was prepared and cleaned.
- **Store Dataset:** Merged to add store-specific metadata (e.g., location).
- **Item Dataset:** Merged to add item-specific metadata (e.g., item categories).
- **Transactions Dataset:** Merged to incorporate the number of transactions per store and date.
- **Oil Dataset:** Merged to provide economic context through oil prices, filling missing values with monthly averages.
- **Holidays Dataset:** Expanded and merged to mark holidays relevant to each store's city.

These steps resulted in a comprehensive dataset combining sales data with store information, item details, transaction trends, economic indicators, and holiday events, enabling a robust foundation for further analysis and modeling. This integrated dataset supports detailed analysis and predictive modeling to understand sales dynamics and forecast future demand.

V. Exploratory Data Analysis

The goal of the exploratory data analysis (EDA) is to understand the structure of the integrated dataset, identify patterns, anomalies, and relationships, and generate hypotheses for further analysis. This phase focuses on understanding key drivers of sales, forecasting demand, and evaluating model performance.

(a) Sales Trends by Day of the Week

- **Observation:** Sales volumes tend to vary by day of the week.
- **Insight:** Weekends (Saturday and Sunday) show higher average unit sales compared to weekdays. This indicates increased shopping activity during weekends.

(b) Monthly Sales Trends

- **Observation:** There are seasonal patterns in the sales data.
- **Insight:** Certain months show consistently higher sales volumes, particularly around holiday seasons. December shows a peak in sales leading up to Christmas, but

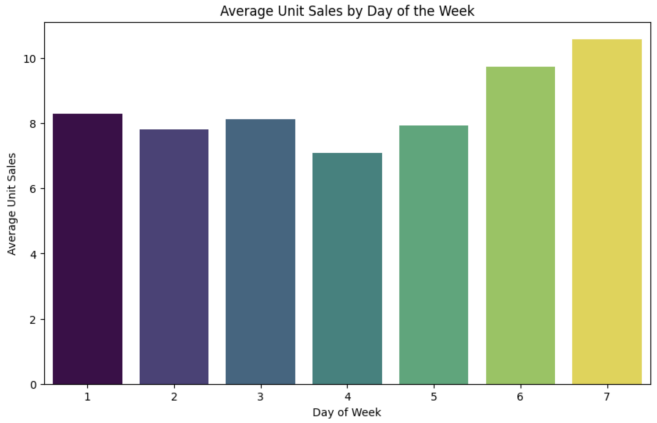

Fig. 1. Sales Trends by Day of the Week

sales drop significantly on December 25th when stores are closed.

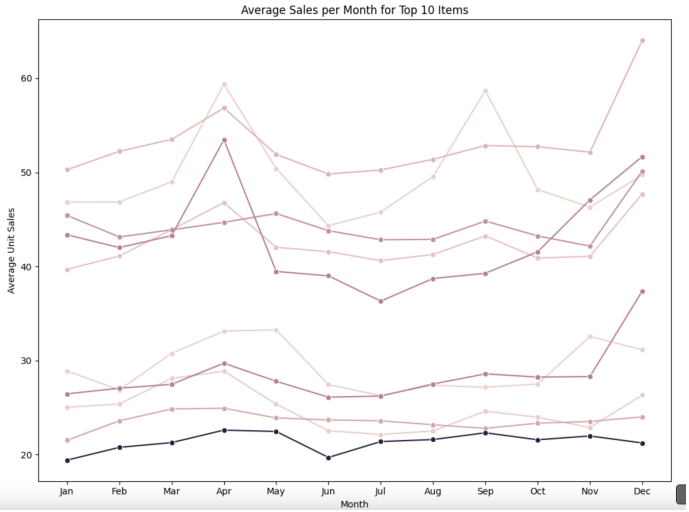

Fig. 2. Monthly Sales Trends

(c) Top-Selling Items

- **Observation:** A small subset of items accounts for a large proportion of total sales.
- **Insight:** The top 50 items (by frequency of sales) contribute significantly to overall sales. These high-performing items are likely staples or popular products across multiple stores.

(d) Store Performance

- **Observation:** Different stores exhibit varying sales volumes and patterns.
- **Insight:** Some stores consistently outperform others, possibly due to location, size, or customer demographics. Analyzing store-specific trends can help in resource allocation and targeted marketing strategies.

(e) Impact of Promotions

- **Observation:** Promotions influence sales volumes.

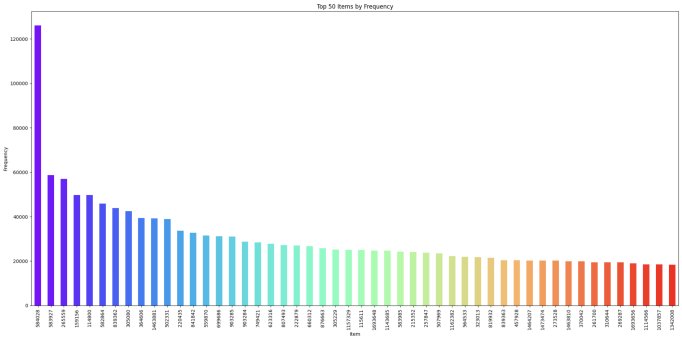

Fig. 3. Top-Selling Items

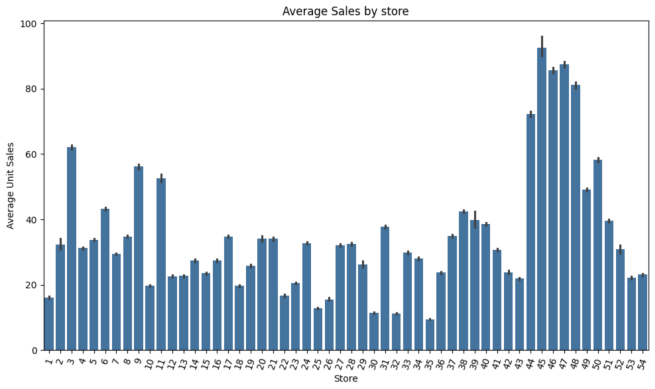

Fig. 4. Store Performance

- **Insight:** Items on promotion show a noticeable spike in sales. Understanding the impact of promotions can aid in optimizing promotional strategies.

(f) Transaction Volume Patterns

- **Observation:** The number of transactions correlates with sales volumes.
- **Insight:** Stores with higher transaction counts typically report higher sales. Transaction peaks occur around major holidays, especially Christmas Eve. Transaction lows are observed around New Year's Day.

(g) Impact of Oil Prices

- **Observation:** Fluctuations in oil prices may correlate with sales trends.
- **Insight:** Higher oil prices may lead to decreased sales due to reduced discretionary spending. Oil price trends could be a proxy for economic conditions affecting consumer behavior.

(h) Holidays and Sales

- **Observation:** Sales are influenced by holidays.
- **Insight:** Significant sales spikes are observed during holidays like Christmas and Black Friday. Regional and local holidays also impact sales, but to a lesser extent than national holidays.

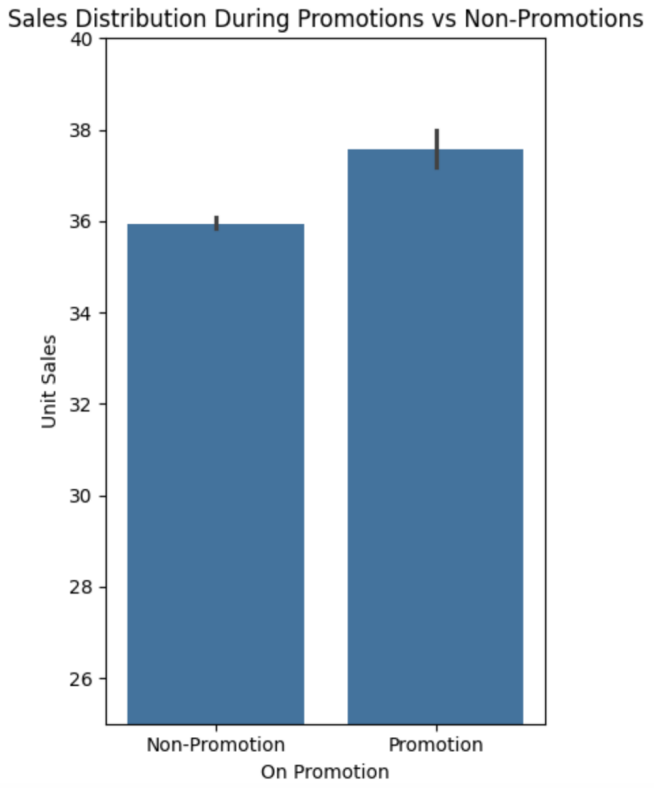

Fig. 5. Impact of Promotions

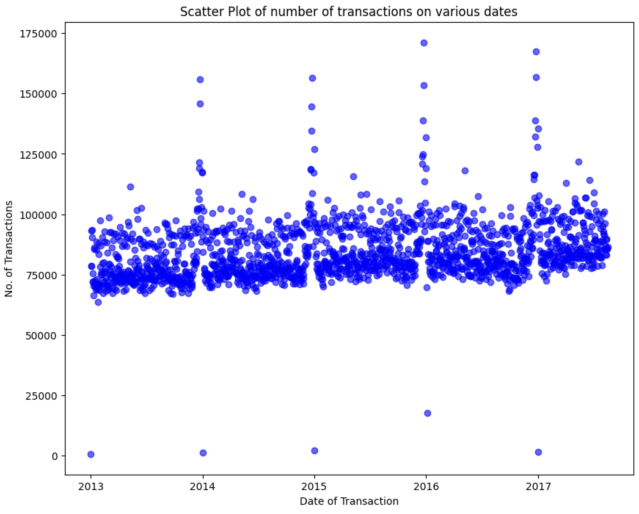

Fig. 6. Transaction Volume Patterns

VI. Observations before Methodology

(a) Overall Sales Patterns

- **Observation:** Sales are driven by both temporal and categorical factors.
- **Insight:**
  - **Temporal Factors:** Day of the week, month, and holidays significantly affect sales.

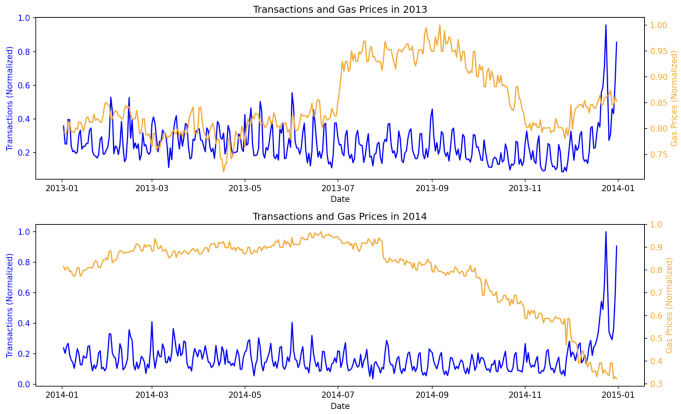

Fig. 7. Impact of Oil Prices

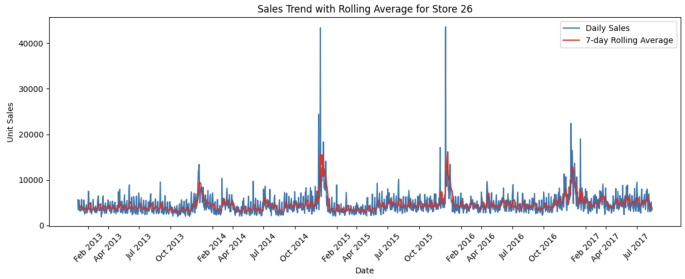

Fig. 8. Holidays and Sales

  - **Categorical Factors:** Item type (e.g., perishable vs. non-perishable), store location, and promotions also drive sales trends.

(b) Data Downsampling

- **Observation:** Due to the large size of the integrated dataset (13GB), analysis was limited to a subset.
- **Insight:** Analysis focused on 10 representative items across all stores. This allowed efficient computation while maintaining meaningful insights.

(c) Seasonal Trends

- **Observation:** Clear seasonality exists in the sales data.
- **Insight:** Sales increase during holiday seasons and decrease in January. Seasonal patterns can be leveraged to forecast demand more accurately.

(d) Economic Indicators

- **Observation:** Oil prices were used as an economic indicator.
- **Insight:** Higher oil prices generally correspond to lower sales, indicating reduced consumer spending during periods of higher economic strain. Incorporating oil price slabs into models helped capture these economic effects.

(e) Holidays Impact Analysis

- **Observation:** Holidays lead to distinct sales patterns.
- **Insight:** Sales increase before major holidays and drop immediately after. Recognizing these patterns helps in

preparing for inventory surges and staffing needs.

VII. Methodology

Model Selection: SARIMA SARIMA (Seasonal AutoRegressive Integrated Moving Average) is a statistical time-series forecasting method that extends the ARIMA model by adding the ability to model seasonality. It is defined by the parameters (p, d, q)(P, D, Q, m) where p is Number of lag observations included (AR – AutoRegressive), d is Number of times the data needs to be differenced to become stationary (I – Integrated), q is Number of moving average terms (MA – Moving Average), P, D, Q are Seasonal counterparts of the ARIMA parameters and m is The number of periods in each season (e.g., 7 for weekly seasonality).

Applications

- **Demand Forecasting:** Predicting product demand based on historical sales data.
- **Retail Analytics:** Forecasting sales trends for stores and products.
- **Economics:** Modeling seasonal trends in economic indicators.
- **Inventory Management:** Estimating future demand to optimize stock levels.

Why SARIMA Fits Our Dataset

- **Time-Series Data:** The sales dataset consists of sales records over time for multiple items and stores, making it a suitable candidate for time-series modeling.
- **Seasonality:** The data shows clear seasonal patterns, such as increased sales during weekends and holidays.
- **Limited Resources:** Given the constraints of computational resources (lack of GPU), SARIMA provides a computationally efficient way to model time-series data compared to complex models like neural networks.

Data Preparation

- **Downsampling the Data:** Due to the large size of the integrated dataset (13GB), analysis was limited to 10 representative items across all stores and a single store to simplify the computation.
- **Train-Test Split:** The data was split into train and test sets based on a chronological order. The first 80% of the time-series data which is time period from 2013 to 2016. The remaining 20%, used to evaluate model performance i.e., data for 2017.
- **Stationarity Check** Ensuring that the time-series data is stationary (constant mean and variance) is crucial for SARIMA. Differencing was applied to remove trends and seasonality.

Model Implementation

- **Model Initialization:** SARIMA was initialized with parameters (p, d, q)(P, D, Q, m), determined through experimentation and analysis of the autocorrelation and partial autocorrelation plots.
- **Parameter Tuning:**
  - **Grid Search:** Different combinations of (p, d, q) and (P, D, Q, m) were tested to find the optimal parameters.
  - **Seasonality:** A weekly seasonality (m = 7) was used to capture patterns occurring over a week.
- **Model Training:** The SARIMA model was trained on the training data. Training involved fitting the model to the historical sales data and learning the underlying temporal patterns.
- **Model Testing:** The model was tested on the unseen test data to evaluate how well it could predict future sales. Forecasts were generated for the test period and compared against the actual sales.

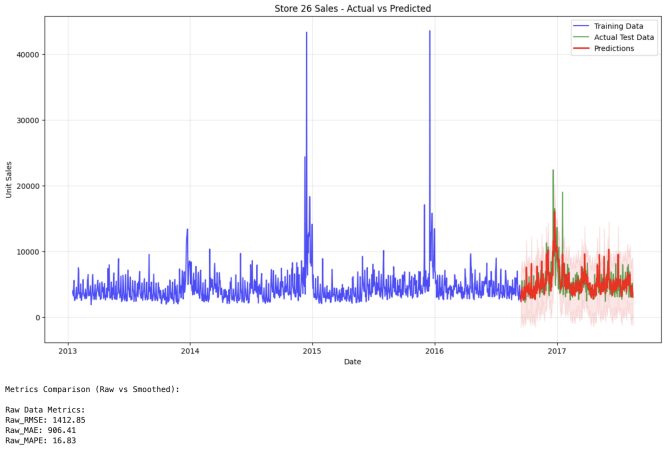

Fig. 9. Results for Store 26

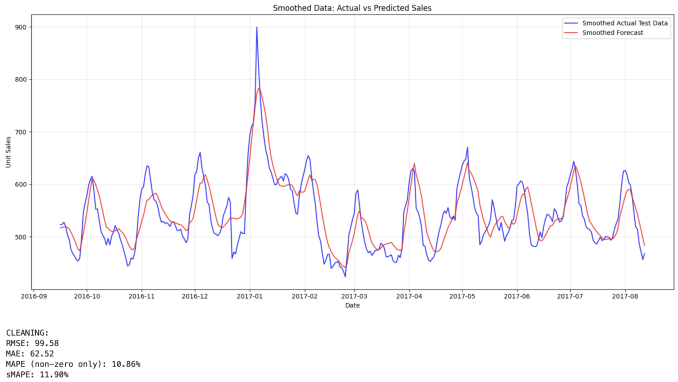

Fig. 10. Results for CLEANING PRODUCT: 168927

VIII. Model Evaluation

To keep our work validated for every dimension we are planning to apply the following methods:

(a) **Mean Absolute Error (MAE):** Measures the average magnitude of errors in predictions.
(b) **Mean Squared Error (MSE):** Penalizes larger errors more heavily compared to MAE.
(c) **Residual Analysis:**Analyzing the residuals (differences between actual and predicted values) to ensure they are random and normally distributed.

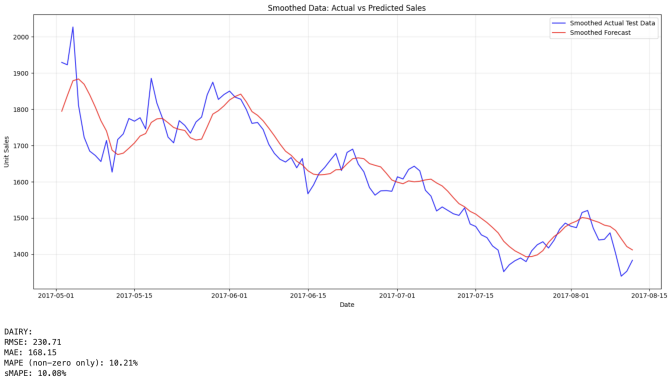

Fig. 11. Results for DAIRY PRODUCT: 1963277

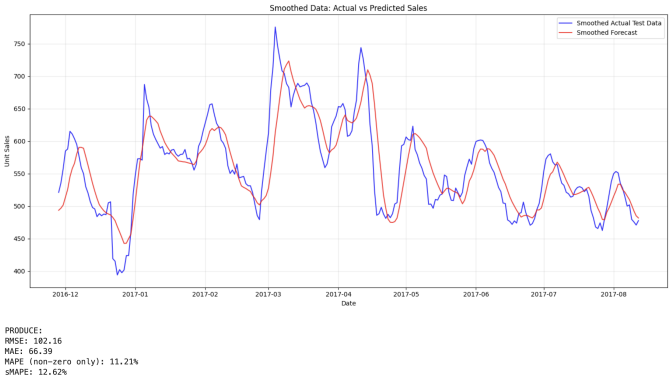

Fig. 12. Results for PRODUCE PRODUCT: 1489837

IX. Future Work

Limitations

(a) **Computational Resources:** Due to limited access to GPUs and other high-performance resources, the analysis was constrained to a small subset of items and a single store.

(b) **Model Complexity:** While effective for simple time-series forecasting, does not account for external factors like economic conditions such as Oil prices, inflation rates, etc. The Competitive Activity like Sales data from competing stores. Also Promotional Campaigns like Marketing efforts influencing sales.

Future Directions

(a) **Neural Networks:** With more computational power, models like Long Short-Term Memory (LSTM) or Temporal Convolutional Networks (TCN) can be used to incorporate multiple factors such as economic indicators, competition, promotions, and holidays also handle complex, non-linear relationships in the data.

(b) **Expanded Data Scope:** Forecasting for all items and all stores rather than limiting to a subset

(c) **Incorporating Exogenous Factors:** Adding variables like oil prices, holiday flags, and transaction volumes to improve forecast accuracy.

Dataset	Smoothed_RMSE	Smoothed_MAE	Smoothed_MAPE
Store_26 Overall Sales	645.64	442.73	7.58%
PRODUCT: 168927	99.58	62.52	11.90%
PRODUCT: 1963277	230.71	168.15	10.08%
PRODUCT: 1489837	102.16	66.39	12.62%

TABLE I
SUMMARY OF THE RESULTS FOR SARIMA MODEL USED IN THIS IMPLEMENTATION.

X. Team Contributions

1) **Teja Swaroop Pothala:** Gathering datasets (sales, stores, items, oil, transactions, holidays). Integrating all six datasets into a cohesive dataset. Identifying seasonality, promotion effects, and anomalies.
2) **Charan kumar Grandam:**Selecting and implementing the SARIMA model. Preparing the integrated dataset. Ensuring the integrity and accuracy of the merged dataset.
3) **Dinesh Devanaboina:** Tuning model parameters and optimizing performance. Proposing future improvements (e.g., neural networks, incorporating external factors). Visualizing trends (e.g., sales by day, month, item, and store). Analyzing the integrated dataset to uncover patterns and insights.
4) **Venkata Sai Sriram Potluri:** Cleaning data, handling missing values, and ensuring consistency. Merging datasets on appropriate keys (e.g., date, store_nbr, item_nbr). Documenting limitations and recommendations for future work.
5) **Laxmi Prasanna Kasireddy:** Conducting residual analysis and interpreting results. Evaluating model performance using metrics like MAE and MSE. Splitting data into train and test sets and training the model.

References

[1] K. Afrin, B. Nepal, and L. Monplaisir, 'A data-driven framework to new product demand prediction: Integrating product differentiation and transfer learning approach', Expert Systems with Applications, vol. 108, pp. 246–257, 2018.

[2] R. Tugay and S. G. Oguducu, 'Demand prediction using machine learning methods and stacked generalization', arXiv preprint arXiv:2009.09756, 2020.

[3] X. Li, Y. Zheng, Z. Zhou, and Z. Zheng, 'Demand prediction, predictive shipping, and product allocation for large-scale e-commerce', Predictive Shipping, and Product Allocation for Large-Scale E-Commerce (March 12, 2019), 2019.

[4] T. Qu, J. H. Zhang, F. T. S. Chan, R. S. Srivastava, M. K. Tiwari, and W.-Y. Park, 'Demand prediction and price optimization for semi-luxury supermarket segment', Computers & industrial engineering, vol. 113, pp. 91–102, 2017.