Online Retail

Project Proposal

Submitted to: Vietnam Datathon

Submitted by: ByteMe (SBS)

Date: 23rd November, 2025

1. Description

This project focuses on **demand forecasting** using the *Online Retail* dataset, which contains all transactions made between 2010 and 2011 by a UK-based online store specializing in selling unique gifts. The dataset includes details such as invoice numbers, product descriptions, quantities, invoice dates, unit prices, and customer information.

By analyzing historical sales data, the project aims to identify **purchasing patterns and seasonal trends**, enabling more accurate forecasts of future product demand. These insights can support better **inventory management**, **marketing strategies**, and **business decision-making** to optimize overall performance and customer satisfaction.

2. Objectives

The objective of this project is to analyze historical sales data from the *Online Retail* dataset (2010–2011) to identify purchasing patterns, trends, and seasonality in customer demand. By applying appropriate forecasting techniques, the project aims to develop an accurate demand prediction model that can estimate future product demand. These forecasts will support effective inventory management, reduce overstocking or stockouts, and enhance marketing and sales strategies. Ultimately, the project seeks to provide data-driven insights that improve business efficiency, decision-making, and customer satisfaction.

3. Outcomes

The expected outcomes of this project include the development of a reliable demand forecasting model capable of accurately predicting future product demand based on historical sales data. The analysis will provide valuable insights into customer purchasing behavior, seasonal trends, and high-demand periods. These outcomes will enable businesses to make informed decisions regarding inventory management, marketing strategies, and resource allocation. Additionally, the project is expected to demonstrate how data-driven forecasting can enhance operational efficiency, minimize waste, and improve overall customer satisfaction.

4. Problem Statement

Retail businesses often struggle with stock management inefficiencies due to unpredictable demand. Overstock leads to storage costs and losses, while understock causes missed sales opportunities and reduced customer satisfaction. The Online Retail dataset contains transaction records from a UK-based online store specializing in gift products from 2010 to 2011. The team aims to analyze this historical data to forecast future demand for various product categories, helping businesses optimize inventory management and supply chain operations. The intended approach addresses the critical business challenge of stockouts and overstock situations that cost businesses millions annually.

5. Literature Review

Demand forecasting has long been a central focus in retail management research, particularly as businesses increasingly depend on data-driven decision-making to optimize inventory, supply chain operations, and customer satisfaction. Numerous studies emphasize that accurate demand prediction helps reduce stockouts, avoid overstocking, and improve operational efficiency. The Online Retail dataset used in this project has been widely referenced in academic research due to its transactional richness, customer-level granularity, and temporal continuity, making it a valuable resource for forecasting models and consumer behavior analysis.

5.1. Retail Demand Forecasting and Inventory Optimization

Retail demand forecasting is fundamentally driven by identifying historical purchase patterns, seasonality, and product-level trends. According to Chopra & Meindl (2016), effective forecasting reduces inventory costs, enhances product availability, and supports strategic planning across the supply chain. In e-commerce settings, where customer demand fluctuates significantly, researchers highlight the importance of fine-grained transactional data to model purchasing behavior (Fildes et al., 2019). The Online Retail dataset, containing more than 500,000 records, allows for this level of analysis, offering detailed insights into purchase quantities, timestamps, prices, and returns.

5.2. Time Series Approaches in Forecasting

Time series models such as ARIMA and SARIMA are among the most commonly used forecasting techniques for retail data. Box & Jenkins (1976) established ARIMA as a foundation for modeling data with autocorrelation, trend, and seasonality. Later studies demonstrate that SARIMA performs well when retail sales exhibit strong seasonal variations - common in gift product sectors, especially around holidays like Christmas and Valentine's Day (Hyndman & Athanasopoulos, 2018). These models provide interpretable outputs and are suitable for datasets where temporal patterns are stable and well-defined.

However, time series models often struggle when faced with irregular patterns, missing values, or non-linear relationships - issues typical in large-scale retail datasets. This challenge has led to the rise of machine learning approaches to complement or outperform traditional statistical methods.

5.3. Machine Learning Models for Retail Forecasting

Machine learning has gained prominence due to its ability to model complex, non-linear interactions among multiple predictors. Studies by Ahmed et al. (2010) showed that tree-based models like Gradient Boosting Machines (e.g., XGBoost) often outperform classical time series models in demand forecasting, especially when incorporating multiple features such as pricing, product category, customer attributes, and historical trend indicators.

XGBoost, as demonstrated by Chen & Guestrin (2016), improves accuracy further through gradient boosting, regularization, and handling sparse data effectively—common in transactional retail datasets.

These algorithms excel when feature engineering is applied, such as deriving temporal features (day, month, holiday tags), customer metrics (RFM analysis), and demand indicators (moving averages and trend decomposition). Such engineered features help machine learning models capture dimensions of retail demand that pure time series models might miss.

5.4. Deep Learning Models for Sequential Retail Data

Deep learning, especially Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, has been widely researched for sequential prediction tasks. LSTMs address limitations of traditional RNNs by capturing long-term dependencies and handling complex temporal dynamics in sales data (Hochreiter & Schmidhuber, 1997). Research by Borovykh et al. (2017) and Bandara et al. (2020) highlights that LSTMs outperform classical models in scenarios with highly fluctuating demand and non-linear trends.

Applying LSTMs to retail forecasting is particularly beneficial when dealing with large transaction volumes and multiple product categories, as these models learn temporal patterns directly from raw sequences rather than relying heavily on manual feature engineering. However, deep learning models typically require more computational resources and careful hyperparameter tuning.

5.5. Ensemble Forecasting and Model Stacking

Ensemble methods are widely considered to yield more robust and accurate predictions compared to individual models. According to Makridakis et al. (2020), ensemble forecasting reduces model-specific bias and exploits the strengths of diverse algorithms. In retail forecasting research, ensembles combining SARIMA, XGBoost, and LSTM frequently deliver superior performance across metrics such as RMSE and MAPE.

Weighted averaging, stacking, and meta-learning approaches provide flexible ways to merge predictive outputs, ensuring stable results even when the underlying dataset includes noise, outliers, or irregular intervals.

5.6. Customer Behavior and RFM Analysis in Demand Forecasting

Several studies emphasize the importance of customer segmentation in improving demand prediction. The RFM (Recency, Frequency, Monetary) model, originally developed for marketing analytics, has been shown by Kumar & Reinartz (2016) to correlate strongly with purchasing likelihood and repeat buying behavior. Integrating customer-level features enhances forecasting models by linking transactional history with future demand patterns.

Given that approximately 25% of customer IDs are missing in the Online Retail dataset, researchers suggest techniques such as anonymized grouping, imputations, or customer-behavior modeling at the invoice level to mitigate incomplete data issues (Borst, 2021).

5.7. Data Quality Challenges in Retail Datasets

Large retail datasets commonly present challenges such as missing data, negative quantities (returns), inconsistent product descriptions, and outliers. According to research from Kotsiantis et al. (2006), model performance significantly depends on the quality of preprocessing, making steps such as deduplication, outlier treatment, and standardization essential. In retail forecasting, proper handling of returns and missing CustomerID values is particularly critical to ensure the reliability of demand predictions.

6. Methodology

This project adopts a structured and data-driven methodology that includes data preprocessing, exploratory analysis, feature engineering, model development, evaluation, and interpretation. The methodology ensures that the forecasting models produced are accurate, reliable, and applicable to real-world retail demand scenarios.

6.1. Research Design

The study follows a quantitative research design, using historical transaction data to build predictive models for future product demand. The methodology integrates statistical analysis, machine learning, and deep learning techniques to compare forecasting capabilities and select the most effective model. A modular workflow enables iterative refinement and reproducibility.

6.2. Data Preprocessing

Data preprocessing is essential due to data quality issues in the Online Retail dataset.

Data Cleaning Steps

- Remove records with missing essential identifiers (InvoiceNo, StockCode, Description).
- Handle missing CustomerID (25% of records) by assigning anonymous identifiers or grouping by invoice.
- Separate or remove negative quantities representing returned items.

- Identify outliers in price and quantity using the Interquartile Range (IQR) method.
- Standardize inconsistent product descriptions using text cleaning and NLP-based normalization.

Data Transformation

- Convert InvoiceDate into usable time formats (daily, weekly, monthly).
- Aggregate product-level demand to required forecasting intervals.
- Convert categorical fields (e.g., StockCode, Country) into numeric/encoded features where necessary.

6.3. Exploratory Data Analysis (EDA)

Exploratory analysis is conducted to understand key sales patterns that influence forecasting:

- **Trend Analysis:** Long-term demand trends across the 2010–2011 timeline.
- **Seasonality Analysis:** Monthly and holiday-driven purchasing patterns.
- **Product Analysis:** Top-selling items, product categories, and lifecycle tendencies.
- **Customer Behavior:** Purchase frequency, recency, spending patterns (RFM metrics).
- **Correlation Analysis:** Relationships between product price, quantity, season, and demand shifts.
- **Geographical Analysis:** Country-wise demand variations (if included).

Visualization tools such as line plots, heatmaps, and distribution charts support the interpretation of data behavior.

6.4. Feature Engineering

Feature engineering enhances model accuracy by incorporating meaningful predictors.

Temporal Features

- Day of week, month, quarter, year
- Holiday indicators (e.g., Christmas season)
- Weeks since last purchase
- Seasonal tags

Customer Features

- RFM (Recency, Frequency, Monetary) metrics

- Customer segmentation
- Customer lifetime value (CLV) indicators

Product Features

- Price elasticity
- Product popularity score
- Cross-selling patterns
- Product lifecycle stage (e.g., introduction, growth)

Demand Indicators

- Rolling averages (7-day, 30-day, 90-day)
- Lag features (t-1, t-7, t-30)
- Trend decomposition components
- Demand volatility metrics

These engineered features allow machine learning models to capture non-linear and complex patterns in retail demand.

6.5. Model Implementation and Validation

- Dataset is split into training, validation, and test sets using time-based splits.
- Models undergo hyperparameter tuning (GridSearch, Bayesian optimization).
- Performance is validated on unseen 2011 data to simulate real-world forecasting.
- Cross-validation ensures model stability across different time windows.

6.6. Deployment and Visualization

After selecting the best-performing model:

- A forecast dashboard prototype is developed using interactive visualizations.
- Plots include time series forecasts, seasonal trends, and product-level predictions.
- The insights support business decision-making, inventory planning, and marketing strategies.

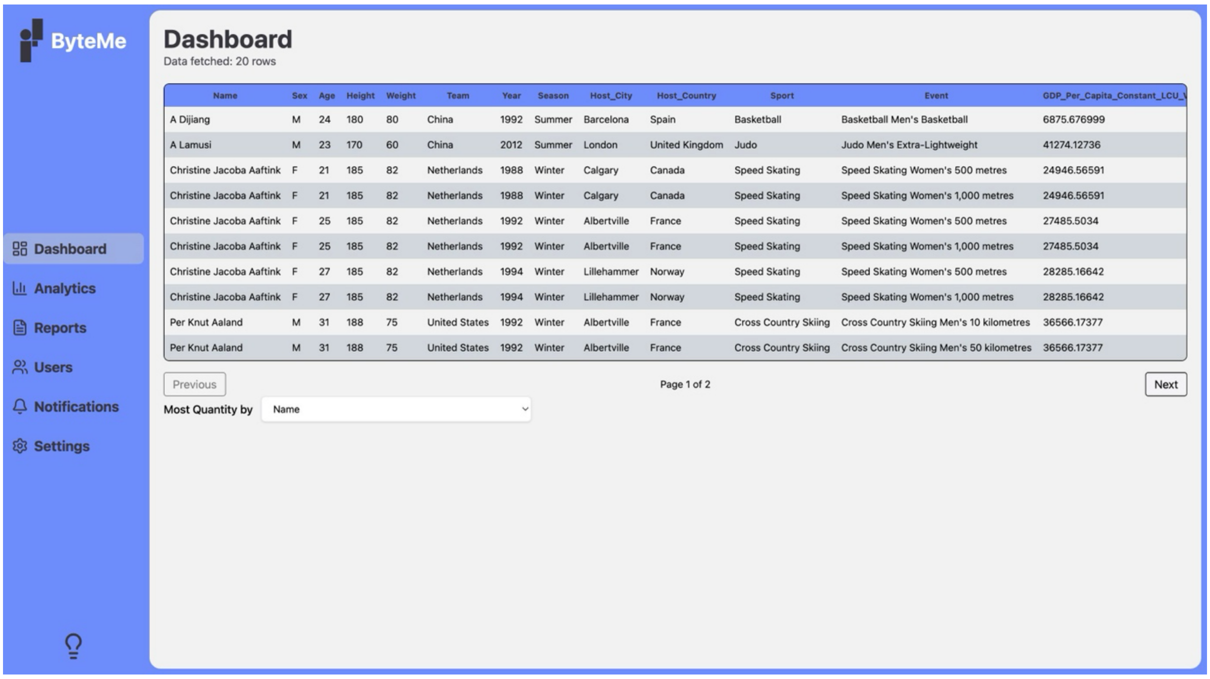
6.7. Ethical and Responsible Use

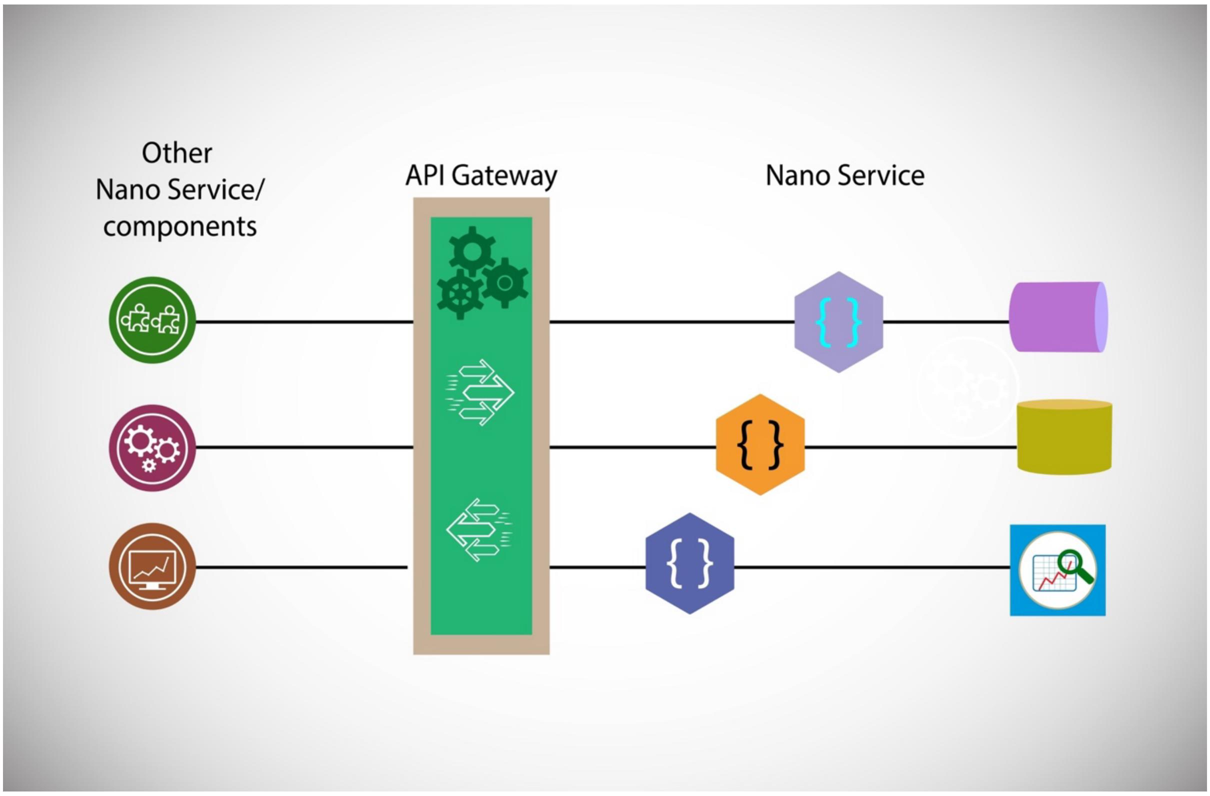
- Customer identities are anonymized to protect privacy.
- Models are monitored for bias, such as overfitting to specific customer groups or seasons.

- Forecast results are used responsibly to support—not replace—human decision-making.

6.8. Visualization

The visualization process is divided into three main phases: **Backend Preparation**, **API Creation**, and **Frontend Development**, following the flow of the MERN architecture.





6.8.1. Backend Preparation (M & N)

This phase focuses on ensuring the final output of your machine learning models is ready to be consumed by the web application.

Step	Component	Clarification	Role
A. Data Export to DB	MongoDB	The Data Analyst (Sitt Min Thar) and Technical Lead will export the final, crucial model outputs into a MongoDB collection. This includes: the final time series forecast values, the actual sales data for comparison, the calculated MAPE/RMSE metrics, and the RFM segment data.	Technical Lead / Data Analyst
B. Server Setup	Node.js	The Visualization Lead (Thu Htet Naing) initializes the Node.js environment. This is the foundation that allows the server-side JavaScript to run and host the Express API.	Visualization Lead

6.8.2. API Creation (E)

This is the bridge that allows the front end (React) to request the data efficiently from the database (MongoDB) and potentially the forecasting model itself.

Step	Component	Clarification	Role
C. Express API Development	Express.js	The Visualization Lead builds an Express server and defines specific API endpoints. Examples: GET /api/forecasts (fetches the main time series data), GET /api/metrics (fetches the model's accuracy KPIs), and GET /api/rfm (fetches customer segmentation data).	Visualization Lead
D. Data Integration	Express.js / MongoDB	The API is coded to connect to the MongoDB collection (Step A). When the React dashboard makes a request to /api/forecasts, Express queries MongoDB and returns the data as a JSON object. This architecture ensures the Python environment used for ML is decoupled from the web application.	Visualization Lead

6.8.3. Frontend Development (R)

This phase focuses on building the user-facing interface and rendering the data into interactive charts.

Step	Component	Clarification	Role
E. Component Design	React.js	The Visualization Lead structures the dashboard into reusable components (e.g., <ForecastChart>, <KpiCard>, <RiskTable>). This follows the principles of good front-end design, making the dashboard easy to maintain and extend.	Visualization Lead
F. API Consumption & State	React.js	Each component uses functions (like fetch or Axios) to call the Express API endpoints (Step C) when the dashboard loads. React manages the retrieved data using state to dynamically update the visualization.	Visualization Lead
G. Interactive Visualization	Plotly.js / Recharts (Libraries)	The fetched data (JSON) is passed to a charting library (e.g., Plotly or Recharts). This library renders the data as interactive plots, such as: zooming time series charts and hover-to-reveal RFM scatter plots, providing the end-user with maximum utility.	Visualization Lead

H. Deployment	MERN Full Stack	The final application is bundled and deployed to a cloud host (like Heroku or Netlify for the front-end, with the API running on an Express server) for the final presentation.	Visualization Lead

7. Team Roles & Responsibilities

7.1. Team Skill Matrix and Cross-Functional Strength (New Section)

To address the key criteria of **Team Strength** and **Skill Gap**, we map our primary and secondary capabilities.

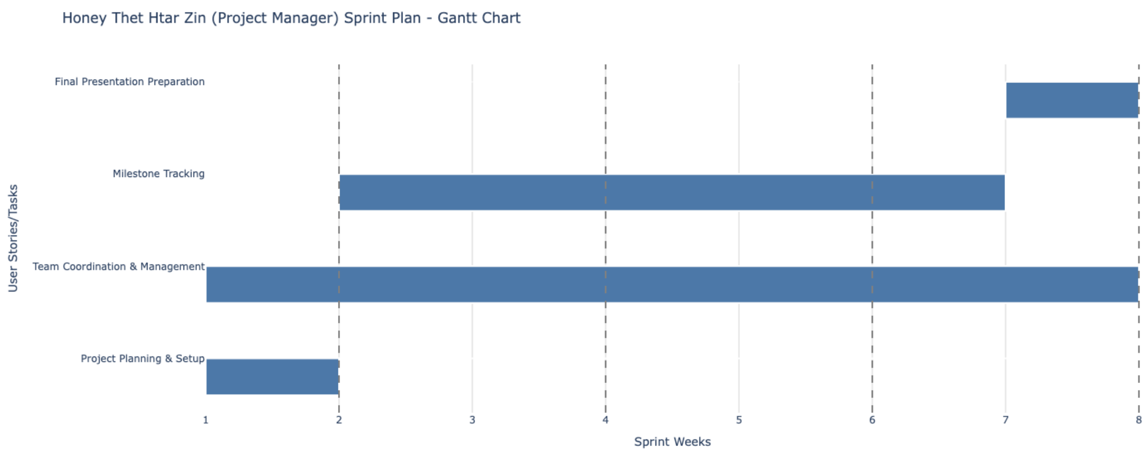
Team Member	Project Management (PM)	Data Analysis (DA)	Python/ML Modeling (ML)	Web Dev (MERN)	Documentation (Doc)	Presentation (Pres)
Honey Thet Htar Zin	P	S	S	-	S	-
Sitt Min Thar	S	P	P	-	-	S
Thu Htet Naing	S	S	S	P	-	-
Thu Kha Kyaw	-	-	-	-	P	S
Phone Myat Min	S	S	-	-	S	P
(P = Primary Skill, S = Secondary/Supporting Skill)						

Skill Gap Mitigation Statement: The team's structure ensures minimal dependency on any single member. For example, if the Data Analyst (Sitt Min Thar) were unavailable, the Project Manager (Honey Thet Htar Zin) and the Web Developer (Thu Htet Naing) have secondary skills in Python and data manipulation (DA, ML) to ensure continuity in preprocessing and model fitting. The Documentation Lead and Presenter also have secondary support roles, guaranteeing project artifacts and external communication are maintained.

7.2. Project Manager - Honey Thet Htar Zin

Honey Thet Htar Zin serves as the **Project Manager**, responsible for overseeing the overall planning, execution, and successful delivery of the project. Key responsibilities include:

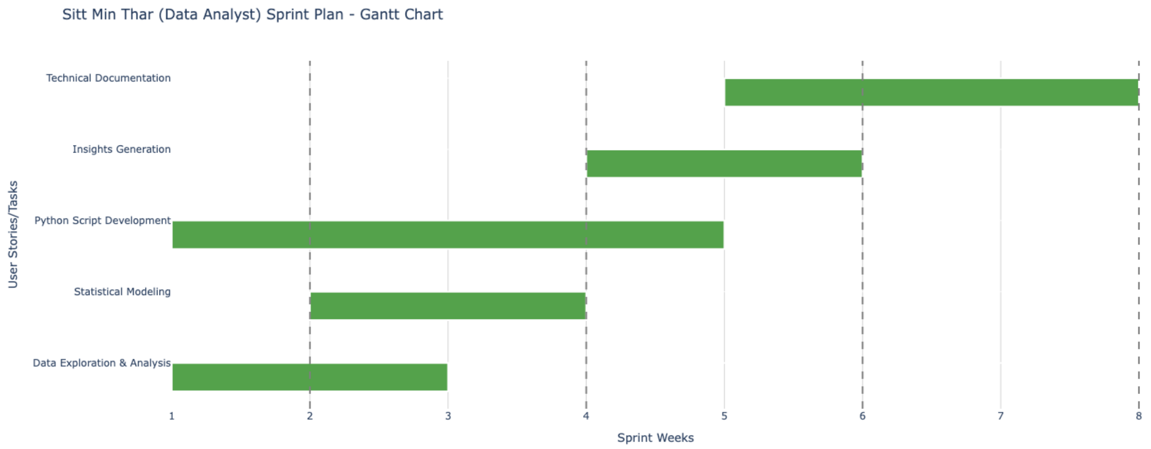
- **Project Planning & Scheduling:** Develops project timelines, defines scope, and allocates resources to ensure all tasks are completed within deadlines.
- **Team Coordination:** Leads and coordinates communication among team members, ensuring everyone understands their tasks and responsibilities.
- **Risk & Issue Management:** Identifies potential risks early, proposes mitigation strategies, and resolves project issues effectively.
- **Quality Assurance:** Ensures all project deliverables meet the required standards and align with project objectives.
- **Stakeholder Communication:** Acts as the primary point of contact between the team and stakeholders, providing regular updates and progress reports.
- **Progress Monitoring:** Tracks project performance, monitors milestones, and adjusts plans when necessary to keep the project on schedule.
- **Decision-Making & Leadership:** Supports team members, makes strategic decisions, and ensures a collaborative and productive working environment.



7.3. Data Analyst – Sitt Min Thar

Sitt Min Thar serves as the **Data Analyst**, responsible for handling, preparing, and analyzing data to support the project's decision-making and system development. Key responsibilities include:

- **Data Collection:** Gathers relevant data from appropriate sources and ensures its accuracy and reliability.
- **Data Cleaning & Preparation:** Organizes and processes raw data, removes inconsistencies, and prepares structured datasets for use by the development team.
- **Data Analysis:** Examines datasets to identify patterns, trends, and important information that support project planning and requirements.
- **Data Validation:** Ensures data accuracy and correctness by checking for errors, duplicates, or missing values.
- **Documentation of Findings:** Records analysis results and data-related observations clearly for team reference.
- **Requirement Support:** Assists in shaping system requirements and logic by providing data-driven insights.
- **Team Collaboration:** Works closely with developers and the project manager to ensure the data aligns with project goals and technical needs.



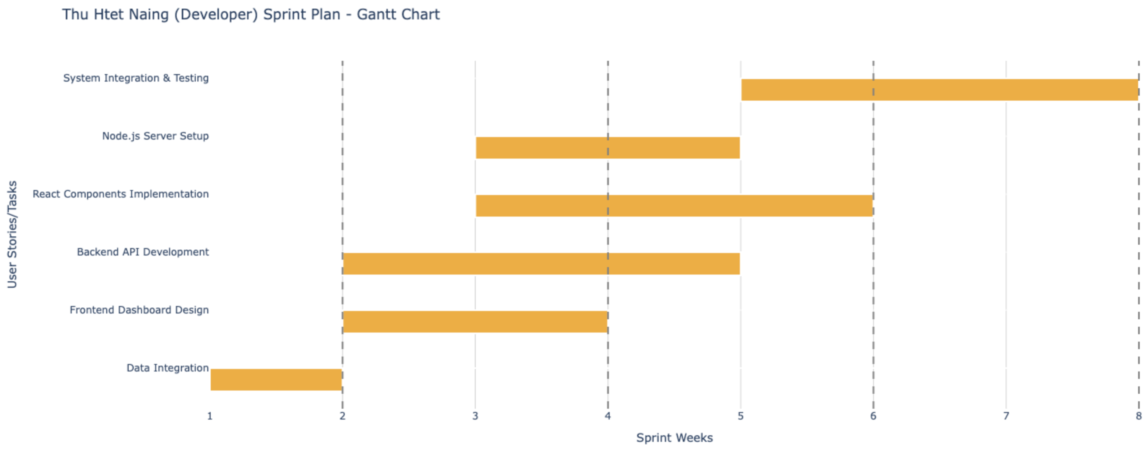
7.4. Visualization & Web Development – Thu Htet Naing

Thu Htet Naing serves as the **Technician**, responsible for developing the web application and handling system visualization tasks. He specializes in **MERN (MongoDB, Express.js, React.js, Node.js) technologies** and supports the project with strong technical implementation skills.

Key responsibilities include:

- **Web Development (MERN Stack):** Builds and maintains core system functionalities using MongoDB, Express.js, React.js, and Node.js.
- **Frontend Implementation:** Develops responsive and user-friendly interfaces to ensure a smooth user experience across devices.
- **Backend Development:** Implements APIs, server-side logic, and database integration to support application workflows.

- **Data Visualization:** Creates interactive and meaningful data visualizations to support dashboards, analytics, and reporting features.
- **System Integration:** Works closely with team members to ensure frontend and backend components operate seamlessly.
- **Technical Support:** Assists in solving technical issues, debugging, and optimizing system performance.
- **Collaboration:** Coordinates with the Project Manager, Data Analyst, and other team members to align technical work with project objectives.
- **Portfolio Reference:** Demonstrates relevant experience and capabilities through his portfolio: https://kukue-portfolio.netlify.app/

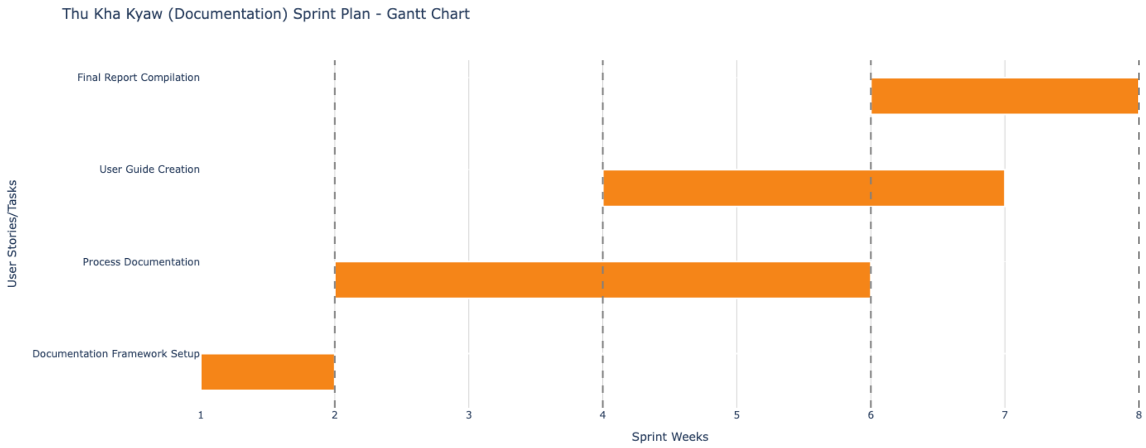


7.5. Documentation Lead – Thu Kha Kyaw

Thu Kha Kyaw serves as the **Documentation Lead**, responsible for preparing, organizing, and maintaining all project-related documents. His role ensures that the project is well-documented, clearly structured, and aligned with academic and project requirements.

Key responsibilities include:

- **Documentation Preparation:** Develops and maintains all project documentation, including project proposals, reports, system descriptions, and user manuals.
- **Information Organization:** Structures content clearly and logically to ensure easy understanding for readers and stakeholders.
- **Version Control:** Updates documents regularly and maintains proper version tracking to reflect the latest project progress.
- **Quality Assurance:** Reviews and edits documents for clarity, grammar, formatting, and consistency with project standards.
- **Collaboration:** Works closely with the Project Manager and Technical Team to ensure all technical information is accurately documented.

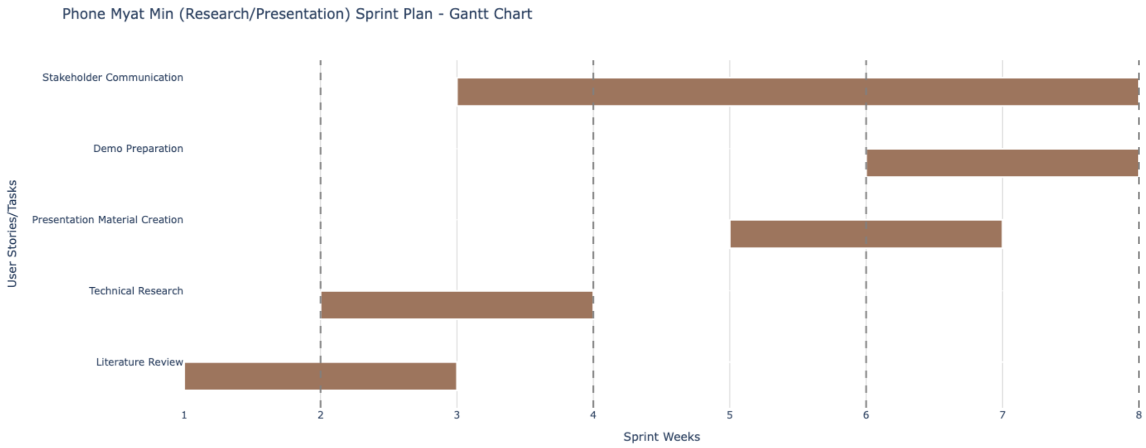


7.6. Technical Presenter – Phone Myat Min

Phone Myat Min serves as the **Technical Presenter**, responsible for delivering clear and engaging presentations about the project's technical components and overall progress.

Key responsibilities include:

- **Presentation Development:** Prepares presentation slides, diagrams, and supporting materials to explain the project's technical workflow and outcomes.
- **Technical Communication:** Simplifies complex technical concepts into understandable explanations for instructors, evaluators, and stakeholders.
- **Project Demonstration:** Leads the live demonstration of system features, functionalities, and user interactions.
- **Coordination:** Works with developers and the Project Manager to ensure all presented information is accurate and up-to-date.
- **Audience Engagement:** Ensures confident and clear delivery during presentations, addressing questions and showcasing the project effectively.



8. Dataset Overview and Analysis

The UCI Online Retail dataset is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Key Features:

- Approximately 541,909 transaction records
- 4,070 unique products (StockCode)
- 4,372 customers (CustomerID)
- Price range from £-11,062.06 to £38,970.00
- Quantity range from -80,995 to 80,995 (negative values indicate returns)

Data Quality Issues Identified:

- 25% of records missing CustomerID
- Negative quantity values indicating returns
- Possible outliers in price and quantity values
- Inconsistent product descriptions

9. Proposed Approach

9.1. Data Analysis and Preprocessing

Data Cleaning:

- Remove records with missing essential fields (InvoiceNo, StockCode, Description)
- Handle negative quantities (returns) by either removing or treating as separate category

- Identify and treat outliers in price and quantity using IQR method
- Standardize product descriptions using NLP techniques

Exploratory Data Analysis:

- Monthly, weekly, and daily transaction patterns
- Customer purchase frequency and value analysis
- Product popularity and seasonal trends
- Geographical analysis (based on UK postal codes)
- Correlation analysis between product categories

9.2. Feature Engineering

Temporal Features:

- Day of week, month, quarter
- Holiday indicators (UK public holidays)
- Days since last purchase
- Seasonal indicators (Christmas, Valentine's Day, etc.)

Customer Behavior Features:

- Recency, Frequency, Monetary (RFM) analysis
- Customer lifetime value metrics
- Purchase pattern consistency
- Product category preferences

Product Features:

- Price elasticity indicators
- Seasonal demand patterns
- Product lifecycle stage
- Cross-selling potential with other products

Demand Indicators:

- Moving averages (7-day, 30-day, 90-day)
- Trend components
- Volatility measures
- Inventory turnover rates

9.3. Forecasting Models

Model 1: Time Series Approaches

SARIMA (Seasonal Autoregressive Integrated Moving Average):

- Capture seasonal patterns in gift product sales
- Account for trend and cyclical components
- Handle autocorrelation in demand data

Model 2: Machine Learning Models

XGBoost Regressor:

- Gradient boosting for high accuracy
- Handle missing values effectively
- Prevent overfitting through regularization

Model 3: Deep Learning Models

LSTM (Long Short-Term Memory):

- Capture long-term dependencies in sequential data
- Model complex temporal patterns
- Handle variable-length sequences

Model 4: Ensemble Methods

- Weighted average of top-performing models
- Stacking approach with meta-learner
- Cross-validation for robust performance estimation

10. EVALUATION METRICS

Model Performance Metrics:

- Root Mean Square Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)
- Symmetric Mean Absolute Percentage Error (SMAPE)

Business Impact Metrics:

- Inventory cost reduction potential
- Stockout probability reduction
- Customer satisfaction improvement
- Revenue increase from optimized stock levels

11. Expected Outcomes and Deliverables

The project's expected outcomes and deliverables include a comprehensive set of technical, business, and visualization components. On the technical side, the project will produce a cleaned and well-documented dataset, an exploratory data analysis report with visual insights, multiple forecasting models with performance comparisons, and an ensemble model designed for optimal demand prediction, along with a model deployment pipeline for future use. From a business perspective, the deliverables include a demand forecasting dashboard prototype, inventory optimization recommendations, insights into seasonal purchasing patterns, customer segmentation based on purchasing behavior, and a detailed product lifecycle analysis report. To support data interpretation, the visualization components will feature interactive time series plots, heatmaps illustrating seasonal trends, visualizations of customer segments, product popularity rankings, and a dashboard presenting forecast accuracy metrics.

12. Technology Stack

Programming Languages:

- Python 3.8+ for data analysis and modeling
- SQL for data querying and management

Libraries and Frameworks:

- Pandas and NumPy for data manipulation
- Scikit-learn for traditional machine learning models
- XGBoost and LightGBM for gradient boosting
- TensorFlow/Keras for deep learning models
- Statsmodels for time series analysis
- Matplotlib, Seaborn, and Plotly for data visualization
- Jupyter Notebooks for interactive analysis

Development Environment:

- Git for version control
- Docker for reproducible environments
- MLflow for experiment tracking

13. SCALABILITY AND BUSINESS APPLICATION

Our solution is designed with scalability in mind:

- Modular architecture allowing easy addition of new data sources
- Cloud-ready deployment pipeline
- Model retraining framework for continuous improvement

- API-based integration with existing business systems

Business Applications:

- Automated inventory replenishment systems
- Dynamic pricing strategies
- Supply chain optimization
- Marketing campaign timing
- New product launch planning

14. ETHICAL CONSIDERATIONS

Data Privacy: Ensuring customer anonymity in all analyses
Fairness: Avoiding bias in demand predictions across customer segments
Transparency: Providing clear explanations for forecasting decisions
Responsible Use: Ensuring predictions are used to improve customer experience

15. List of Key Personnel

Honey Thet Htar Zin (SBS24060044)	: Project Manager
Thu Htet Naing (SBS25010150)	: Technician
Sitt Min Thar (SBS25010149)	: Data Analyst
Thu Kha Kyaw (SBS24010103)	: Documentation Lead
Phone Myat Min (SBS24010097)	: Technical Presenter

16. Contact Information

For further discussion or clarification regarding this project, please contact:
Name: Mrs. Honey Thet Htar Zin
Student ID: (SBS24060044)
Email: honey.zin0044@sbsuni.edu.mm.