Department of Computer Science and Artificial Intelligence
University of Jeddah
Kingdom of Saudi Arabia

Sales Forecasting using AI

Students

Abdulaziz Alamri 2041681

Abdulrahman Khaled 2040226

Ahmed Munir 2040937

Supervisor

Dr. Mohammed Hamed

Table of Contents

General Introduction

The primary objective of this project is to create and implement a specialized sales prediction software tailored for the retail industry. Accurate sales forecasting is of utmost importance for retailers because it helps them fine-tune inventory management, production planning, and distribution processes. By utilizing historical sales records, customer demographics, product seasonality, and market trends, the software aims to generate dependable sales projections.

In the world of retail, having precise sales forecasts is a key factor for success. It allows retailers to make well-informed decisions about inventory quantities, pricing strategies, and resource distribution. With accurate sales estimates, retail businesses can prevent stock shortages and decrease surplus inventory, leading to cost reductions and enhanced customer satisfaction.

This software will employ cutting-edge data analysis methods and visual representation tools to offer valuable insights into the demand patterns for various products. These insights will help retailers better understand customer preferences, recognize potential growth areas, and optimize their product selection. Furthermore, the software will produce detailed reports that summarize the sales predictions and emphasize important trends and patterns.

By adopting this sales prediction software, retail companies can refine their decision-making processes and secure a competitive advantage in the marketplace. The model in our study achieved an outstanding accuracy of 97.90%, forecasting retail sales growth with unprecedented precision, comparable to analyst revenue forecasts and significantly improving accuracy when combined with analyst forecasts [1]. The software's dependable sales projections will empower retailers to plan their operations more effectively, fulfill customer needs efficiently, and fully realize their revenue potential.

In summary, the goal of this project is to provide retail businesses with a trustworthy sales prediction solution that promotes data-driven decision-making, enhances operational efficiency, and ultimately contributes to their long-term success in the ever-changing and competitive retail landscape.

1. Chapter 1 Introduction

In Chapter 1, we will start by introducing the project and what it's all about. We'll talk about the specific problem we're trying to solve. Our main aim is to achieve certain goals and outcomes through this project. We'll also discuss the objectives of our project, which means the specific things we'll be looking into and studying. We'll get an idea of what to expect in the rest of the report as we outline its structure and organization. Finally, we'll touch upon the project plan, giving you an overview of the timeline and activities, we'll be doing.

1.1 Scope & and the problem to solve.

Forecasting sales in the retail sector is one of the critical elements for the success of companies in this sector. Forecasting aims to identify and predict patterns and trends in product demand, helping companies make informed strategic decisions about inventory management, resource allocation, and pricing strategies.

When a company has historical sales data, this data can be used to analyze past models and develop predictive models for future sales. This includes analyzing influencing factors such as seasons, special events, holidays, and market trends. Based on these analytics, the utilization of seasonal sales forecasting methods is widespread [2]. companies can generate accurate forecasts of future sales.

Sales forecasting helps retail companies improve their operations and increase their profitability. With accurate forecasts, companies can better balance inventory levels. Production planning and supply chain coordination can also be improved based on forecasts, which reduces costs and enhances operational efficiency.

In short, forecasting sales in the retail sector is an essential tool for planning and making strategic decisions. By analyzing historical data and using. With AI and probability technologies, companies can generate accurate forecasts that improve inventory management, save money, and achieve their business goals.

1.2 Problem definitions

In the retail sector, merchants face numerous challenges in predicting future sales for their products. Success and maximizing inventory utilization, production planning, and distribution rely on accurately forecasting upcoming demand. If demand is estimated inaccurately, it can lead to high costs due to excess inventory.

1.2.1 Real Live Example

One of the main reasons for the company's downfall was poor sales forecasting. Toys R Us had been using the same forecasting model for many years, but the retail landscape had changed significantly during that time. E-commerce had become more popular, and consumers were increasingly shopping online for toys. As a result, Toys R Us was not able to accurately forecast demand for its products, and it ended up with too much inventory in some areas and not enough in others. This led to lost sales, increased costs, and eventually bankruptcy [3].

1.3 Aims

- **Develop an accurate sales forecasting model:** Retailers will benefit from having a greater understanding of consumer demand and the ability to make wiser choices regarding inventory, price, and marketing.
- **Provide valuable insights into demand rates for goods**: To manage inventory levels and enhance customer service, this will assist retailers in identifying trends and patterns in demand.
- **Enhance inventory management:** This will assist merchants in reducing excess inventory and stockouts, which can help them save money and increase customer happiness.
- **Optimize production planning and distribution processes:** This will assist merchants in reducing excess inventory and stockouts, which can help them save money and increase customer happiness.
- **Enable retailers to make informed decisions regarding pricing, promotions, and marketing strategies:** Retailers will be able to increase sales and profits as a result.
- **Improve customer satisfaction:** This will support retailers' efforts to draw in and keep customers, which may result in more revenue and earnings.
- **Increase overall profitability and competitiveness of retail businesses:** Retailers can use this to keep one step ahead of the competition and accomplish their corporate objectives.

1.4 Objectives

- **Gather and process relevant data on sales and customer behavior**: This includes collecting data on past sales, customer demographics, and customer behavior.4 This data can be gathered from a variety of sources, such as sales records, customer surveys, and social media data.
- **Identify key variables that have an impact on sales:** This includes identifying factors such as product popularity, price, competition, and economic conditions that can affect sales.
- **Select and implement appropriate forecasting algorithms:** There are a variety of forecasting algorithms available, each with its own strengths and weaknesses. The appropriate algorithm for a particular business will depend on the type of data available, the level of accuracy required, and the time horizon for the forecast.
- **Test the accuracy of the forecasting models against actual sales data**: Once a forecasting model has been selected, it is important to test its accuracy against historical sales data. This will help to ensure that the model is generating accurate forecasts.
- **Generate highly accurate sales predictions based on the input variables:** Once the forecasting model has been tested and validated, it can be used to generate highly accurate sales predictions. These predictions can be used to make decisions about inventory management, staffing, budgeting, and resource allocation.

1.5 Report Outline

- **Introduction**: Importance of sales forecasting in the retail sector and its impact on business success.
- **Literature Review:** Comprehensive review of existing literature on sales forecasting methods and their application in the retail industry.
- **Methodology**: Detailed description of the methodology employed in the project, including data collection, preprocessing, selection of appropriate forecasting algorithms, and validation techniques.
- **Results**: Presentation and analysis of the results obtained from different forecasting models, highlighting their accuracy and effectiveness in predicting sales.
- **Discussion**: In-depth discussion of the implications and insights derived from applying the forecasting model to retail stores, including its impact on inventory management, staffing, budgeting, and resource allocation.
- **Conclusion:** Summary of the project's findings and contributions to the field of sales forecasting in the retail sector, emphasizing its significance and potential for improving business performance.

1.6 Project plan

The project aims to forecast the sales volume of a selected product over a specified period, the plan is structured into several steps: -

Steps	tasks	weeks
Step 1	Data collection and filtering	**W1**
	Selection of suitable algorithms for analysis	
	Division of work among team members.	
Step 2	Initiation of work on initial objectives.	**W2**
	Review of work and reassessment of the plan	**W3**
Step 3	Evaluation of the employed algorithms.	**W4**
	Verification of progress towards initial objectives.	
	Transition to the core project objectives.	**W5**
	Comprehensive project review session for error correction.	**W6**
Step 4	Compilation and verification of final objectives.	**W7**
	Final project presentation.	**W8**
	Project completion and finalization	

1.7 Conclusion

In conclusion, Chapter 1 has provided a comprehensive overview of the project, laying the foundation for the subsequent chapters. The introduction has set the stage by outlining the specific problem addressed in this project – the crucial need for accurate sales forecasting in the retail sector. The scope and importance of forecasting in optimizing various aspects of retail operations, from inventory management to production planning, have been highlighted.

The real-life example of Toys R Us serves as a poignant illustration of the repercussions of poor sales forecasting in the retail industry. This emphasizes the significance of our project's objectives, which are aimed at developing an accurate sales forecasting model to benefit retailers in making informed decisions.

The aims and objectives of the project have been clearly articulated, focusing on the development of a robust forecasting model and the enhancement of inventory management, production planning, and overall decision-making processes. The outlined objectives serve as a roadmap for the subsequent chapters, guiding the research towards practical and tangible outcomes.

Furthermore, the report's structure and organization have been briefly outlined in the report outline section, providing readers with a roadmap for navigating the subsequent chapters. The project plan, with its defined steps and tasks, establishes a clear timeline for the execution of the project, ensuring a systematic and organized approach.

Chapter 1, therefore, serves as a crucial introductory component, offering readers a comprehensive understanding of the project's context, objectives, and significance. As we progress into the subsequent chapters, we will delve into the literature review, methodology, results, and discussion, ultimately providing a thorough analysis of the project's findings and their implications for the field of sales forecasting in the retail sector.

2. Chapter 2 Literature Review

Sales forecasting play vital rule in decision making and strategic planning for businesses operating in retail sectors. accurate sales forecast empowers effective inventory management, resource allocation, improve customer service and needs. The literature review aims to explore existing research and insights on sales forecasting. This section provides an overview of the key finding from selected research papers in the same field, highlighting their contributions and implications for the project.

2.1 Stakeholders definition

- **Business Owners and Managers** - They are the primary stakeholders as they are responsible for making strategic decisions related to sales forecasting, such as inventory management, staffing, and resource allocation. They would be interested in the accuracy of the sales forecasting model and its ability to provide insights into consumer demand trends.
- **Sales and Marketing Teams** - They are responsible for executing the sales and marketing strategies based on the forecasted sales data. They would be interested in the granularity of the sales data, such as product-specific demand and customer segmentation.
- **IT and Data Science Teams** - They are responsible for developing and maintaining the AI model used for sales forecasting. They would be interested in the data collection and preprocessing steps and testing the accuracy of the model.
- **Customers** - They indirectly benefit from the project's outcomes as it enables businesses to provide better customer service and product availability. They would be interested in the impact of the sales forecasting model on store operations and inventory management.

2.2 Project Domain

Our project focuses on sales forecasting in the retail sector, utilizing historical data from stores to predict future sales. The primary objective of this project is to enhance financial returns in the retail industry, forecasting of sales forms the foundation of maximizing profits and successful planning operation at the start of the fiscal year [4], by mitigating losses attributed to stagnant sales and maximizing profits for products with high demand.

Within the retail sector, accurate sales forecasting plays a crucial role in strategic decision-making. It enables companies to identify and predict patterns and trends in product demand, facilitating informed decisions regarding inventory management, resource allocation, and pricing strategies.

By analyzing historical sales data and considering influential factors such as seasons, special events, holidays, and market trends, companies can generate precise forecasts for future sales. The implementation of sales forecasting models empowers retail companies to optimize their operations and enhance profitability.

Accurate sales forecasts enable better inventory management, ensuring optimal stock levels and reducing the risk of overstocking or stockouts. Moreover, these forecasts assist in improving production planning and coordinating the supply chain, resulting in cost savings, and increased operational efficiency.

In conclusion, the area of concentration for our study is sales forecasting for retail businesses. We seek to assist stores to optimize their inventory management, reduce losses, and boost revenue by accurately predicting future sales using historical sales data, advanced analytical methods, and machine learning models.

2.3 Literatures Review

The literature review explores three studies in sales forecasting within the retail sector. Each study adopts unique methodologies, addressing specific challenges and providing valuable insights into the complexities of predicting sales in diverse retail contexts.

2.3.1 Export sales forecasting using artificial intelligence [5].

The research highlights the importance and significant impact of sales forecasting on various levels. Additionally, the research points out the problem of irregular sales, which creates greater difficulty in the forecasting process. The research also indicates the effect of certain variables on subsequent outcomes.

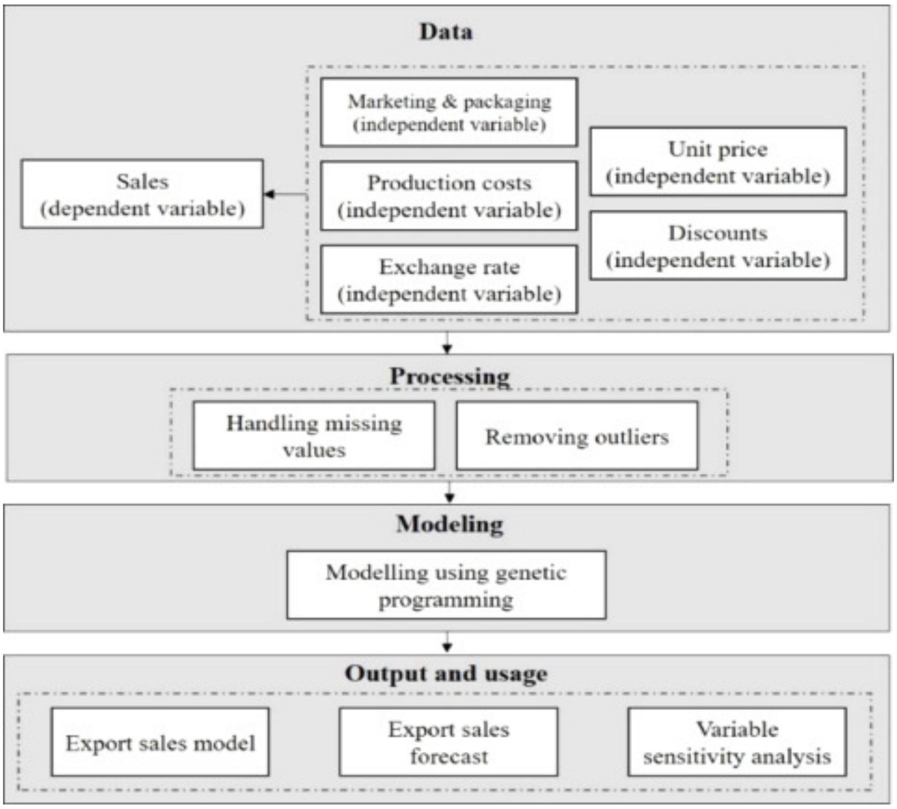


fig 2.3.1.1

The text outlines a four-step process:

1. Data collection: collecting six variables (five independent, one dependent) from an export sales company.

2. Data preprocessing: Addressing missing values and outliers in the data.

3. Modeling: Using a genetic program to develop a new causal forecasting model and evaluating its accuracy.

4. Output and usage: presenting the model, accuracy indices, forecasting export sales for six weeks, comparing with real sales data, and analyzing the impact of each variable on sales.

In summary, the process involves collecting data, preparing it, developing a forecasting model, and presenting the results.

Variable	Sensitivity	% Positive	Positive magnitude	% Negative	Negative magnitude
c	0.9597800	100%	0.9597800	00%	0
m	0.1378100	66%	0.1543600	34%	0.10564
p	0.0081528	100%	0.0081528	00%	0
x	0.0039138	100%	0.0039138	00%	0
d	0	00%	0	00%	0

Key:

- Price (p)
- Demand rate (x)
- Costs (c)
- Discounts (d)
- Packaging and marketing (m)

The sensitivity analysis results show that production costs have the strongest influence on export sales. Higher production costs are associated with higher sales, as they indicate better quality of raw materials and the final product. Marketing and packaging (variable "m") also have a positive impact on sales, but to a lesser extent. Unit price and exchange rate have a consistently positive effect on sales, as they increase the purchasing power of foreign buyers. Discounts, on the other hand, have a negligible impact on sales and can be reduced to allocate resources to other areas like marketing and packaging.

Conclusions

Sales forecasting is crucial for supply chain and production management as it impacts various aspects of an organization. Traditional forecasting techniques may be insufficient when sales behavior becomes unstable or is affected by complex economic factors. This study utilized Genetic Programming (GP) to model export sales for a Middle Eastern company facing fluctuating sales. The GP-based model demonstrated high accuracy and precision, as evidenced by error metrics and the forecast for a six-week period. Variable sensitivity analysis indicated the positive impact of marketing, exchange rate, price, and costs on sales, while discounts had negligible effects. This research addresses the gap in causal export sales forecasting using GP and real empirical data, providing a practical framework for accurate sales forecasting, inventory management, and cost control. The study's applicability extends to other export companies in stable markets, as well as organizations with larger datasets and different regions. Future studies should consider larger datasets encompassing longer sales periods and additional influential variables.

2.3.2 Predicting Sales Revenue by Using Artificial Neural Network in Grocery Retailing Industry: A Case Study in Turkey [6].

This study focuses on the importance and difficulty of sales forecasting and its significant impact on stores and their profits. The study specifically examines the influence of Artificial Neural Networks (ANN) on the accuracy of sales forecasts and compares the results across major stores in Turkey.

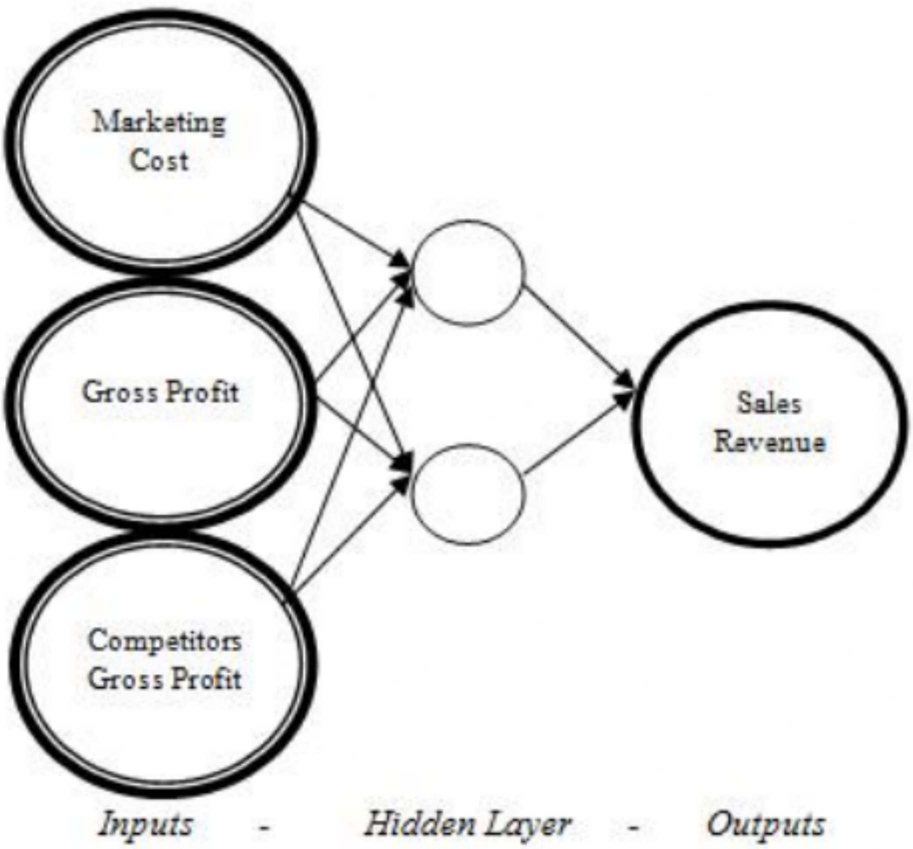


Fig. 1. Neural network of predicting sales revenue.

To address the absence of a linear relationship between variables, two neurons are implemented in one hidden layer of the model. Backpropagation is the preferred method for training in this model. For training purposes, the model is limited to 3000 epochs, and the minimum weight delta is set to its default value of 0.000001. For prediction, the initial weight and learning rate are set at 0.30, while the momentum is set at 0.60. These specific values are chosen because they yield better results. The hyperbolic tangent function is used as it can accommodate both positive and negative data inputs.

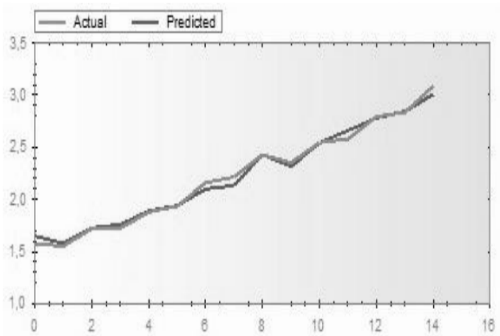


Fig. 2. Bım’s Actual and Predıcted Sales Revenue (Bıllıon Tl vs. Tıme Perıod).

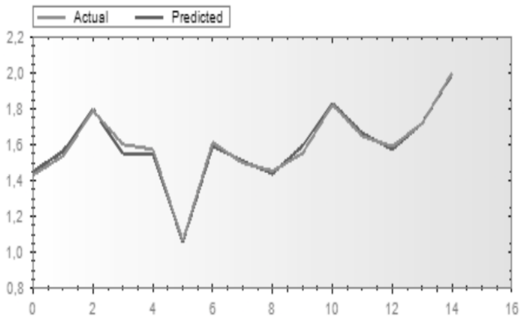


Fig. 4. Mıgros’s Actual and Predıcted Sales Revenue (Bıllıon Tl. vs. Tıme Perıod).

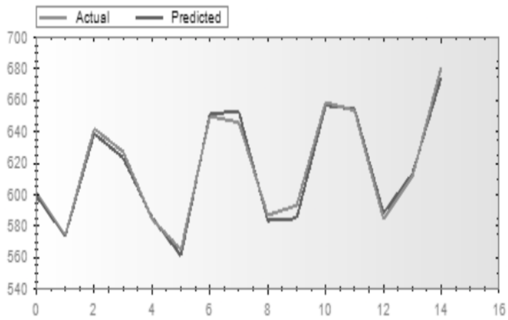


Fig. 3. Carrefour’s Actual and Predıcted Sales Revenue (Mıllıon Tl vs. Tıme Perıod).

Migros is the largest multi-factor retailer in Turkey, generating 7 billion in sales revenue in 2013. Despite a steady increase in gross profit, their sales revenue has not followed the same trend due to rising cost per unit sold. Carrefour, another major player in the Turkish grocery retail industry, had sales revenue that remained relatively stable over the years due to intense competition. The forecasted sales revenue for each retailer showed strong similarities to actual sales revenue, indicating that artificial neural networks (ANN) are suitable for forecasting in the grocery retail industry. Based on constant marketing costs, gross profit, and competitors' gross profit, the forecasted sales revenues for the first quarter of 2014 are as follows: BİM (3,098,697,398 TL), Carrefour (720,487,615 TL), and Migros (1,848,710,093 TL)

Conclusion

Sales forecasting is crucial for every company, especially the big ones. This process is very complex because there are a lot of factors that should be considered. To implement reachable goals and successfully achieve them, companies are keen on predicting the next period's sales. Compared to other methods, ANNs are organic, and this method builds a learning algorithm to predict results better. In this study, sales revenue forecasts are very close to actual sales revenues for each firm. Other factors that could affect sales revenue can also be added to the mix through further research. Also in this study, only the grocery retailing industry is analyzed. To get the big picture, other industries can be considered for analysis in the future.

2.3.3 Daily retail demand forecasting using machine learning with emphasis on calendric special days [7].

The main points addressed in the research can be summarized as follows:

1. Demand forecasting is crucial for retailers in making operational decisions, especially for perishable goods with high deterioration rates.

2. Special calendar days, including public holidays and other events, pose a challenge for demand forecasting due to different demand patterns.

3. The research focuses on a bakery chain, addressing the problem of forecasting daily demand for different product categories at the store level.

4. Machine learning methods are considered as a potential solution for demand forecasting and are compared with traditional approaches.

5. The possibility of formulating the forecasting problem as a classification task is explored.

6. The study evaluates the performance of various machine learning methods, such as artificial neural networks and gradient-boosted decision trees.

7. Machine learning methods are found to be superior to established approaches, and classification-based approaches outperform regression-based ones.

8. Machine learning methods are not only more accurate but also suitable for large-scale demand forecasting scenarios in the retail industry.

9. The research contributes to the existing literature by presenting a real-world application in the bakery domain and discussing the formulation of time series forecasting as a machine learning problem. The empirical evaluation assesses the performance of machine learning methods in predicting sales for different types of special days.

10. The results highlight the importance of considering external factors, such as calendar events, in demand forecasting.

Conclusion

The research focuses on demand forecasting in the retail sector, specifically for special days. It presents a case study of a bakery chain and evaluates the use of machine learning methods for forecasting daily demand at the store level. The study concludes that machine learning methods, particularly classification-based approaches, outperform traditional methods and provide more accurate and suitable forecasts. It highlights the importance of automation and the availability of large datasets in demand forecasting.

2.4 Comparison criteria definition

- **Accuracy**: The degree of similarity between the expected sales data and the actual sales data.
- **Scalability**: The quality of performance with increasing data quantity and number of variables.
- **Interpretability**: The ability to provide insights into the factors influencing sales with detailed explanations.
- **Efficiency**: The ability to predict sales accurately with minimal resource consumption, ensuring efficient resource utilization while maintaining high prediction accuracy.
- **Ease of use**: The ease of inputting data and interpreting results.
- **Cost**: The cost of implementing and maintaining the model.
- **Speed**: The time taken to provide accurate sales forecasts, from data processing to presenting results

2.5 Comparison results and the feasibility

After evaluating the four models, genetic programming, linear regression, random forest, and decision tree, based on the defined criteria of accuracy, scalability, interpretability, efficiency, ease of use, cost, and speed, The results showed that genetic programming outperformed the other models in terms of accuracy and interpretability. However, linear regression was the most efficient model in terms of resource consumption. Random forest and decision tree models were found to be scalable and easy to use, more models may be used in the future.

Criteria	Sales Forecasting	L1	L2	L3
Dataset	Walmart	Export sales company	Grocery retailing industry	-
Applicability	Retail industry	Retail industry	Retail industry	Retail industry
Model	XGBoost Genetic & ANN	Genetic program-based model	Artificial Neural Network	-

2.6 Conclusion

The literature review demonstrates the importance of sales forecasting in retail businesses, highlighting the risks associated with inaccurate predictions. The use of AI and machine learning techniques, such as genetic programming, linear regression, random forest, decision trees, and artificial neural networks (ANNs), offers promising avenues for improving forecasting accuracy and interpretability. The selected research papers provide insights into the practical applications of these models, including cost reduction, inventory management, and decision-making support.

The objective of this project is to develop a sales forecasting model that utilizes the knowledge acquired from the literature review. The model will incorporate the advantages of different techniques, ensuring accuracy, interpretability, and efficiency in predicting sales. Comparing the outcomes of various methods will guide the selection of the most appropriate model for the retail domain's specific demands. This will allow businesses to make informed decisions, optimize resource utilization, and achieve financial objectives.

3. Chapter 3 General Analysis

Chapter 3 provides a comprehensive overview of the sales forecasting project's requirements. It begins by detailing the dataset from Walmart, outlining variables, and conducting a quality assessment. The data cleaning process is explained, followed by insightful explorations of dataset trends. The chapter then outlines functional requirements, specifying key system components and capabilities, along with non-functional requirements emphasizing scalability, accuracy, security, and usability. Hardware requirements are also briefly discussed, setting the groundwork for the subsequent phases of the project.

3.1 Requirements gathering [8].

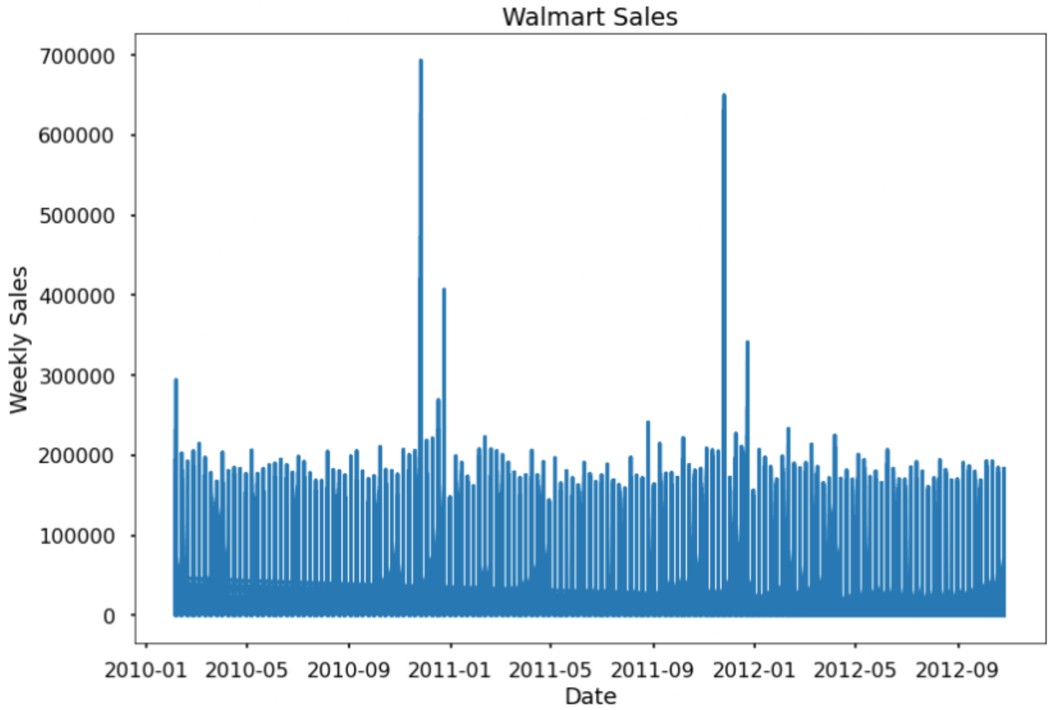
We are using dataset given by Walmart a huge retail stores company, to get insight on the features of the dataset we create this data dictionary: -

Variable	Data Type	Description	Quality Assessment
Store	Integer	The ID of the store	No missing values or errors
Date	Date	The date of the sale	No missing values or errors
Department	Integer	The ID of the department	No missing values or errors
Weekly_Sales	Float	The weekly sales for the department	Missing exist, which may be errors
CPI	Float	The Consumer Price Index for the week	Missing exist, which may be errors
Temperature	Float	The average temperature for the week	No missing values or errors
Unemployment	Float	The unemployment rate for the week	Missing exist, which may be errors
Fuel_Price	Float	The average price of fuel for the week	No missing values or errors
IsHoliday_x	Boolean	Whether the date is a holiday	No missing values or errors
IsHoliday_y	Boolean	Whether the date is a holiday	No missing values or errors

During our meticulous data examination process, we detected instances of missing values within the dataset. To address this, specific strategies were employed, focusing on maintaining the overall quality and integrity of the data [9]. For instance, within the 'features' dataset, missing values in key columns like 'CPI,' 'Unemployment,' and 'Weekly_Sales' were systematically replaced with the corresponding median values. This careful substitution not only preserved the completeness of the dataset but also contributed to its overall high data quality.

Following the essential cleaning procedures applied to the dataset, which included the removal of duplicate features and replacement of missing values with the median to ensure data integrity, our exploration delved into revealing key relationships within the dataset.

The temporal scope of our data spans from February 5, 2010, to October 26, 2012. Among the dataset's attributes, we identified 45 stores and 81 departments, noting that the department composition varies across stores. Stores are categorized into three types, denoted as A, B, and C, based on their sizes.

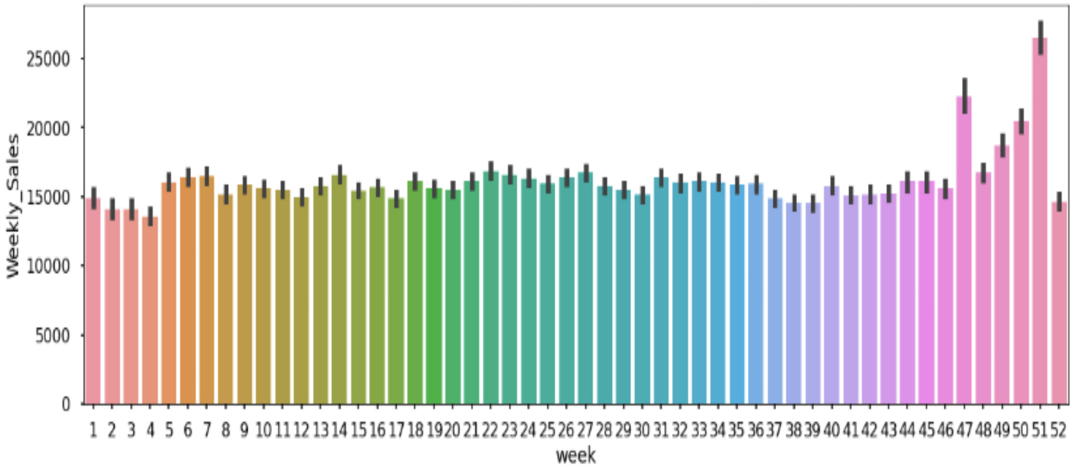


The above line chart tells us that sales have been increasing over time and reach their highest in the holiday season (Nov and Dec), while attaining their lowest during the summer (Jul and Aug).

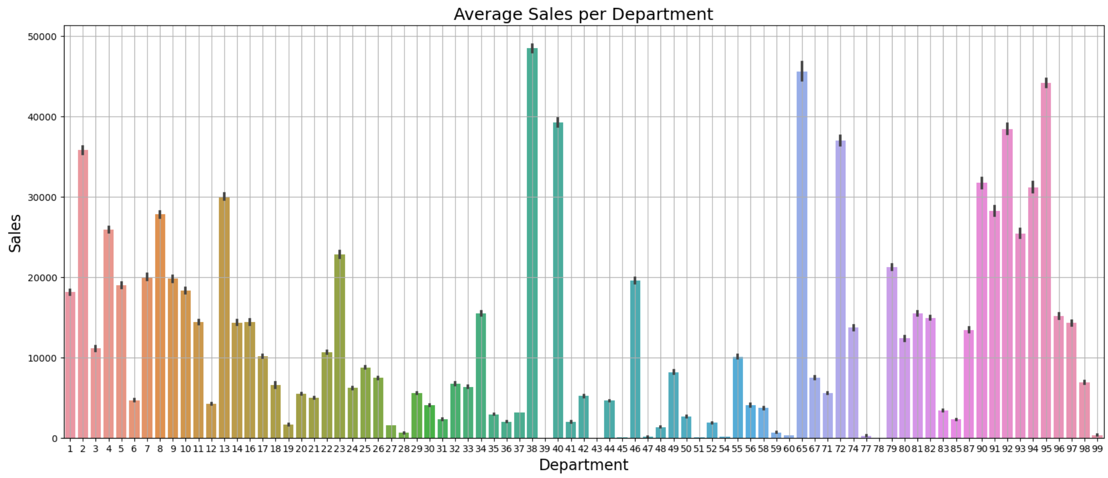
A noteworthy observation is the higher average sales during holiday periods compared to regular dates. Analyzing sales trends over the years, we observed that the sales volume in 2010 surpassed those in 2011 and 2012, with the caveat that data for November and December 2012 is not included.

Additionally, January consistently exhibited lower sales compared to other months, attributed to the post-holiday decline in consumer spending.

Exploring potential correlations, we found that features such as CPI, temperature, unemployment rate, and fuel price may exhibit no discernible or concealed patterns concerning weekly sales. This initial data exploration sets the stage for more in-depth analyses and modeling to derive actionable insights for our sales forecasting system.

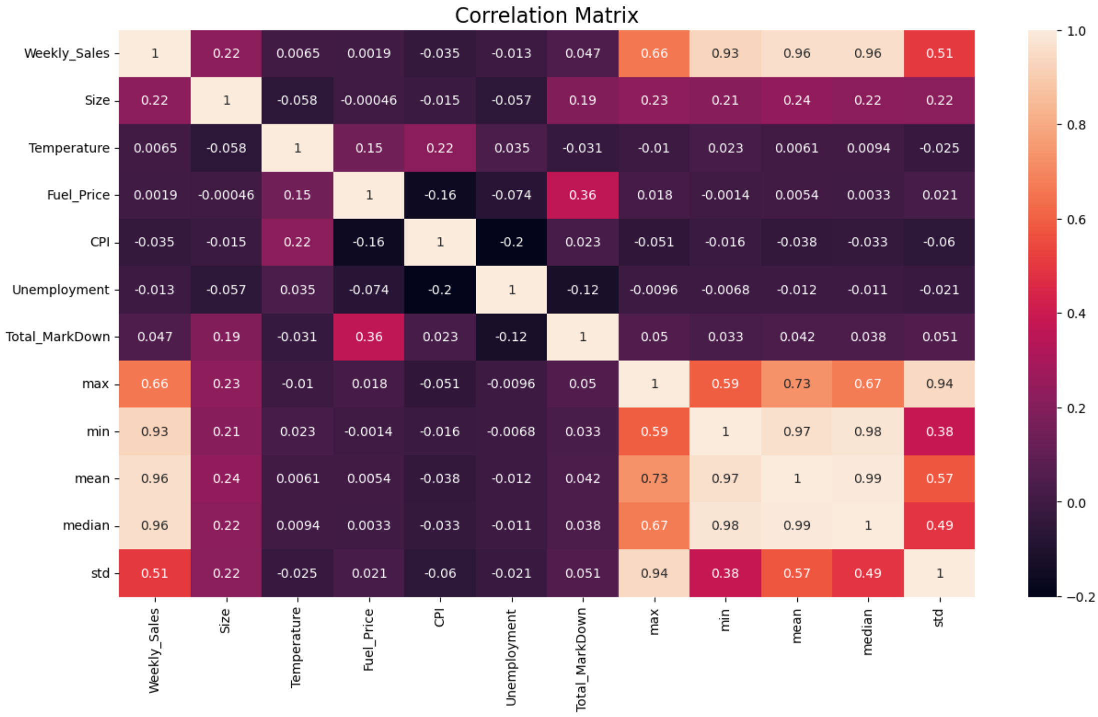


From the analysis of bar graphs, it is evident that the 51st week, corresponding to the Christmas season, and the 47th week, associated with Thanksgiving and Black Friday, exhibit significantly higher average sales. Additionally, when considering the monthly distribution, it becomes apparent that the initial month of the year and the early weeks consistently demonstrate lower sales figures.
This observation may be attributed to various factors, including post-holiday lulls and reduced consumer spending at the beginning of the year. The detailed exploration of monthly and weekly patterns provides valuable insights into the temporal dynamics of sales, allowing for a more nuanced understanding of the underlying trends.



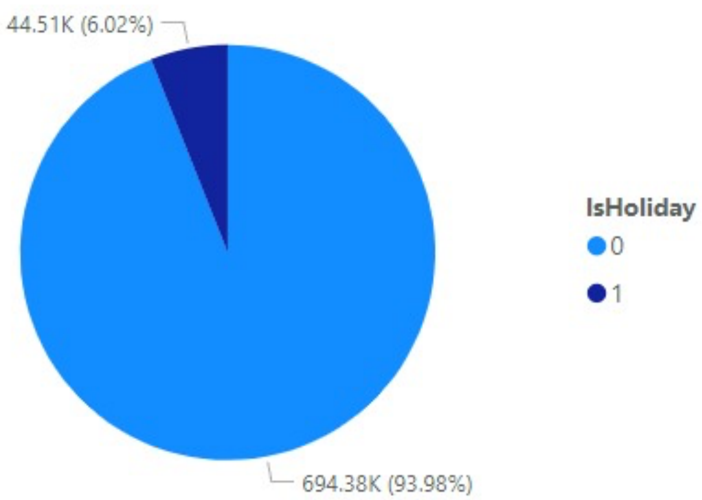
The bar graph illustrating average sales by department provides a quick overview of sales distribution across different segments of our retail stores. Notably, there are departments with zero sales, which is a key observation that warrants further investigation.

Identifying departments with zero sales is crucial for ensuring data integrity and could prompt a closer examination of factors contributing to this phenomenon. This insight is vital for refining our sales forecasting system and addressing any anomalies or challenges associated with specific departments.



Examining the plot, we can identify the strength and direction of the correlations between various features. Strong correlations (closer to 1 or -1) indicate a robust relationship, while weaker correlations (closer to 0) suggest a more modest or negligible connection [10].

This analysis is crucial for feature selection and model building. Identifying highly correlated features helps in avoiding multicollinearity, a condition where two or more variables in a regression model are highly correlated, making it challenging to discern their individual effects on the target variable.



The 'Holiday Distribution' chart serves as a quantitative representation delineating the distribution of weeks categorized as holidays and non-holidays within the dataset. The pie chart reveals that the predominant majority, precisely 93.88%, corresponds to weeks without designated holidays.

In contrast, a comparatively minor fraction, amounting to 6.12%, is attributed to weeks marked as holidays. This graphical depiction offers a precise and insightful overview of the temporal distribution of holidays, providing valuable context for the analysis of sales patterns in relation to holiday occurrences.

3.2 Functional Requirements

- **Data Collection**: The system should collect sales-related data from retail stores by import it to the system, including historical sales records, customer behavior data, and product information.
- **Data Processing:** The system should process the collected data to clean and transform it into a suitable format for analysis, including removing duplicates, handling missing values, and standardizing the data.
- **Variable Identification**: The system should identify key variables that impact sales, such as product type, pricing, promotional offers, seasonal factors, and customer data, also removing unrelated variables.
- **Algorithms Models**: The system should execute appropriate forecasting algorithms, such as time series analysis, regression models, or machine learning techniques, to predict future sales based on the identified variables.
- **Model Training and Validation:** The system should train forecasting models using historical data and validate their accuracy by comparing the predicted 1sales with actual sales data.

- **Sales Forecasting:** The system should generate accurate sales forecasts based on the input variables and trained forecasting models.
- **Integration:** The system should have the ability to integrate with other existing systems and databases to access additional data sources or share generated sales forecasts with relevant stakeholders.

3.3 Non-Functional Requirements

- **Scalability**: The system should be able to handle increasing amounts of data as the business grows without compromising on performance.
- **Accuracy:** The system should provide accurate sales forecasts with a low error rate to ensure reliable decision-making.
- **Security**: The system should have appropriate security measures in place to protect sensitive sales data, including user authentication and access control.
- **Usability**: The system should be user-friendly and intuitive, with clear and concise reporting that enables stakeholders to easily understand the sales forecasts and make informed decisions

3.4 Hardware Requirements

Computer Power: Sales forecasting usually involves processing large amounts of data through complex computation. Therefore, a robust CPU is favorable to handling the calculations efficiently. Also, if applicable to use deep learning, and GPUs can speed up the process [11].

Memory (RAM): Adequate memory is essential when it comes to the loading and processing of extensive datasets. We are considering having at least 16 GB or more of RAM to ensure smooth and efficient processing.

Storage: Sufficient storage is required to store the sales data or any intermediate or results. We are considering using solid-state drives (SSDs) for quick data access.

Cloud Computing: Alternatively, to the above requirements we can use Amazon Web Service (AWS), Microsoft Azure, and Google Collab these platforms offer flexible and scalable computing resources, with varying CPU, RAM, and GPU configuration as well as data storage.

3.5 Conclusion

Chapter 3 serves as a pivotal juncture in our exploration of the sales forecasting project, providing a comprehensive overview of the project's requirements. The chapter commences with a meticulous examination of the dataset procured from Walmart, elucidating key variables and conducting a rigorous quality assessment.

During this scrutiny, instances of missing values are identified, prompting strategic interventions to ensure data integrity. Notably, the careful substitution of missing values with corresponding median values preserves the overall quality of the dataset.

Following the essential data cleaning procedures, the chapter embarks on an insightful exploration of dataset trends. The temporal scope of the data, spanning from February 5, 2010, to October 26, 2012, is outlined. Significant observations, such as the impact of holiday seasons on sales and variations in sales volume across different store sizes, lay the groundwork for subsequent analyses.

Functional requirements are then meticulously outlined, delineating the key components and capabilities essential for the sales forecasting system. From data collection and processing to algorithm execution and integration, each requirement is meticulously detailed, forming a roadmap for the system's development. Non-functional requirements, encompassing scalability, accuracy, security, and usability, are also underscored, emphasizing the system's robustness and user-friendliness.

The chapter delves into hardware requirements, addressing crucial considerations such as computer power, memory, storage, and the potential integration of cloud computing. These specifications are crucial for ensuring the system's efficiency in processing extensive datasets and complex computations.

In conclusion, Chapter 3 not only lays a solid foundation for subsequent phases of the project but also provides a roadmap for the development of an effective sales forecasting system. The insights gained from the dataset exploration, coupled with the detailed requirements, set the stage for informed decision-making and the successful implementation of the sales forecasting project.

4. Chapter 4 Sprint 1 Design & Development

During the initial phase of Sprint 1 Design and Development, our primary emphasis is on establishing the groundwork for our sales forecasting system. This stage encompasses diverse elements, with a notable focus on use cases that visually delineate the interactions among various system actors and the functionalities they necessitate.

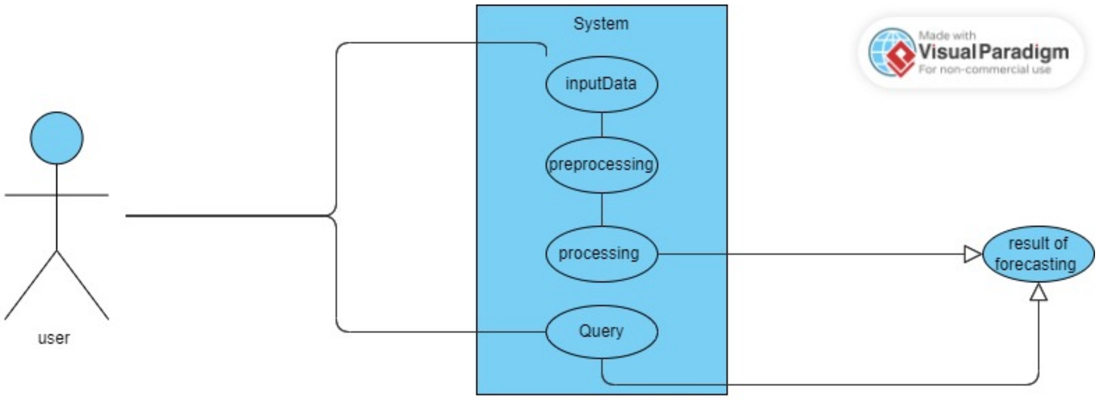
Progressing to the static facet of design entails the creation of the system's structure and the relationships among its components, encompassing classes, modules, and databases. Subsequently, we delve into the dynamic design, facilitating an understanding of how distinct components interact and respond to user actions or system events.

In parallel, we address the architectural facet of design, elucidating the overarching structure and configuration of our system. This stage significantly contributes to defining the holistic framework within which our system operates.

Ultimately, the implementation phase involves the translation of design specifications into executable code. Throughout this process, testing and evaluation play a pivotal role in guaranteeing the quality and reliability of our system.

4.1 Sprint1 Use Case Diagram

The use case diagram represents the key elements and interactions of the sales forecasting project, involving various actors and their roles:



1. User: The user represents an actor interacting with the system. The user initiates a query and provides input data to the system.

2. System: The system represents the software being used It consists of many components: input data and preprocessing and processing.

3. Input Data Preprocessing: This component is responsible for handling and preparing the input data provided by the user before it can be processed further. It may involve cleaning, transforming, or formatting the data to ensure it is suitable for analysis.

4. Processing: This component performs the main processing task, which is likely to be demand forecasting based on the context of the research discussed earlier. It takes the preprocessed input data and applies forecasting algorithms or models to generate the result of the forecasting process.

5. Result of Forecasting: This represents the Query of user and the output or outcome of the forecasting process. It could be the predicted demand values, trends, or any other relevant information related to demand forecasting.

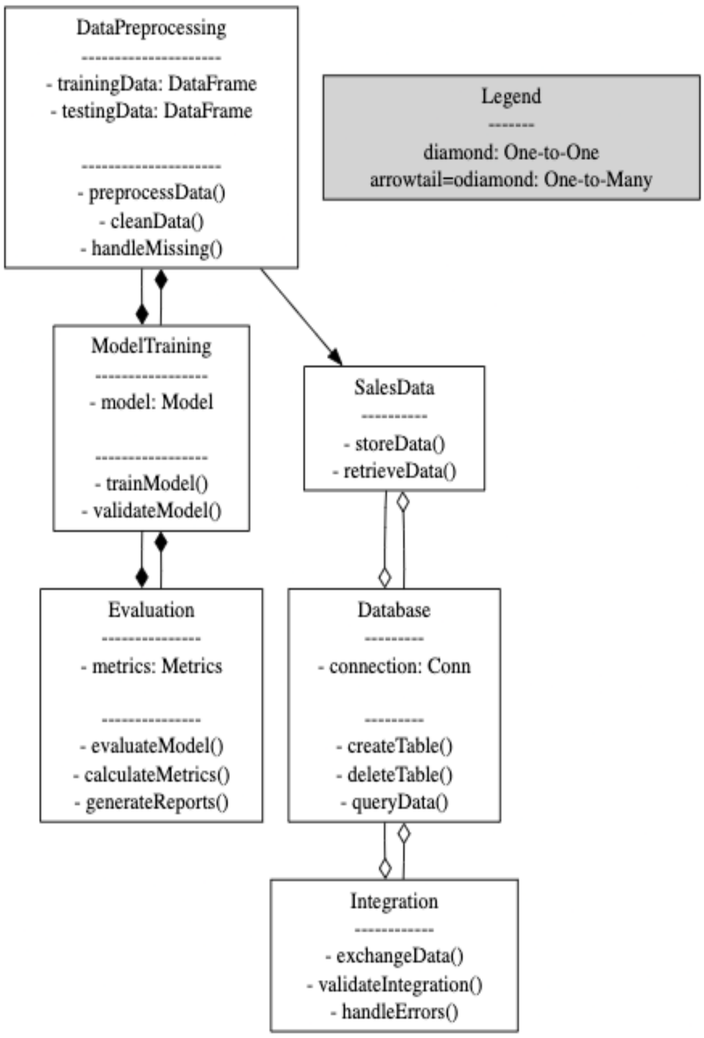
The diagram captures the collaboration between the program developer and the system, where the developer's expertise and input drive the development of the software, while the system's capability to generate forecasts is based on the data and algorithms.

4.2 Sprint 1 Aspect Design

System Engineering Design:
Design the system architecture that supports sales forecasting using machine learning Identify the required components for data preprocessing, model training, and evaluation.Explore appropriate frameworks and libraries for machine learning tasks.

Data Storage Design:
Design the database structure to store sales data.Define tables and relationships for storing training and testing data.Ensure data integrity, efficient storage, and retrieval.

Integration Design:
Specify mechanisms for data exchange between systems.Ensure consistent and reliable data integration.

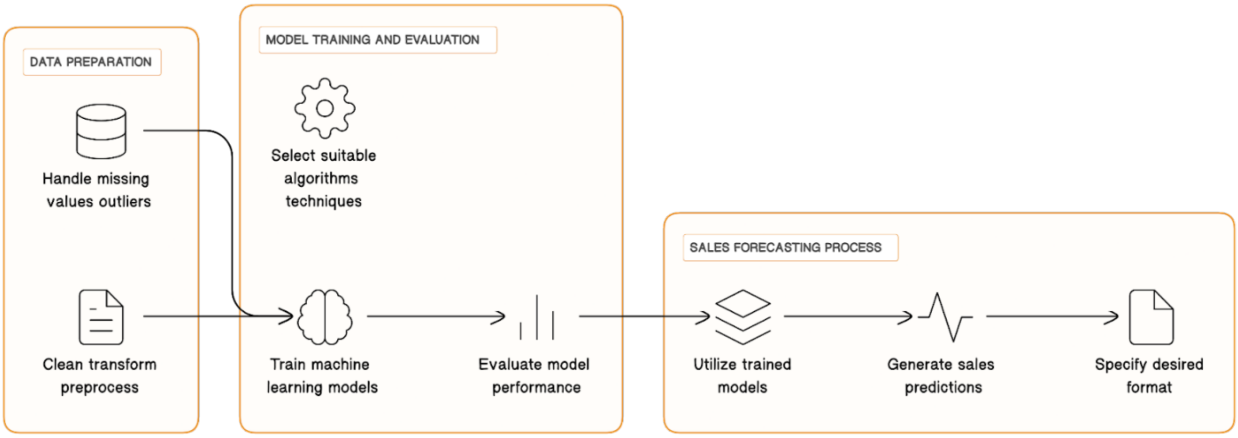


4.3 Sprint 1 Dynamic Aspect Design

Data Preparation: Process the existing data for machine learning tasks.
Clean, transform, and preprocess the data appropriately.
Handle missing values and outliers effectively.

Model Training and Evaluation:
Train machine learning models using the available data.
Select suitable algorithms and techniques for sales forecasting.
Evaluate the model performance using the testing data.

Sales Forecasting Process:
Utilize the trained models to generate sales predictions.
Provide inputs such as historical sales data and relevant variables.
Specify the desired format for presenting sales forecasts.



4.4 Sprint 1 Architectural Aspect Design

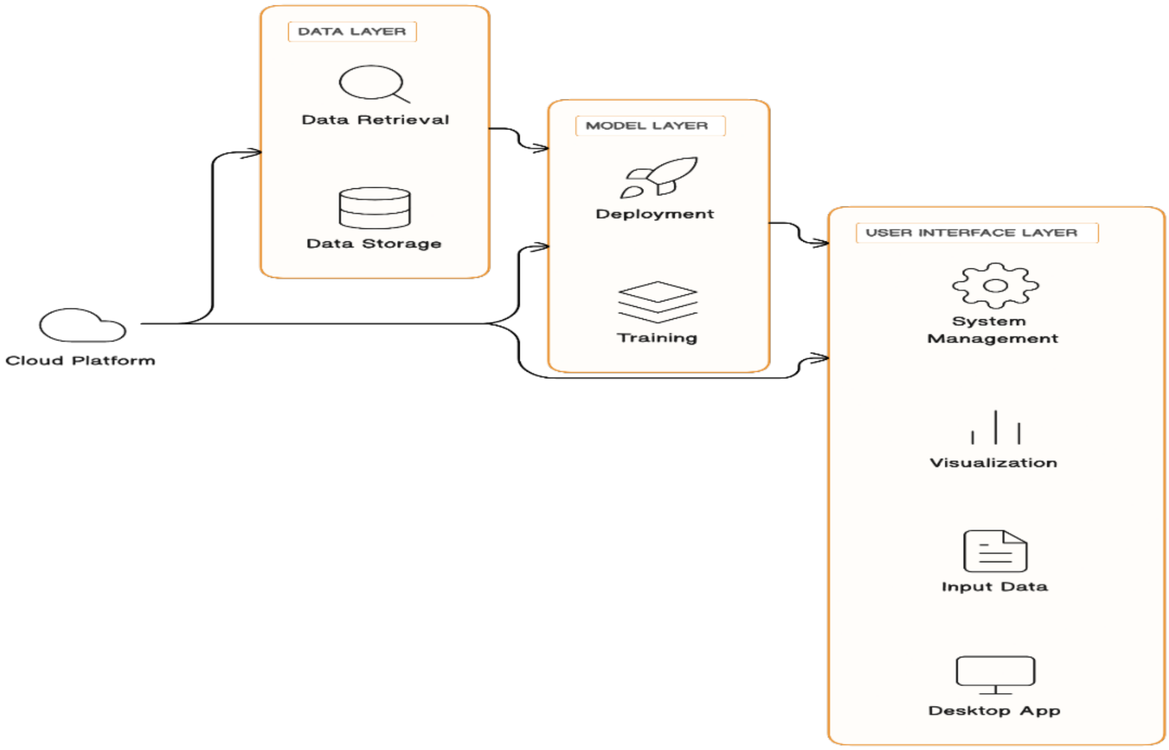
in this section we outline the architectural aspect design of our project, the system will include three main components, the data layer, the model layer and the user interface layer, with the support of cloud-based platforms we will ensure scalability and flexibility in meeting user demands.

- **Data Layer:** the data layer serves as the base layer of the system, which will oversee storage and retrieval of data.

- **Model Layer:** the model layer is the core of the system, training and deploying various machine learning models.

- **User Interface Layer:** this layer acts as the getaway for interaction between users and the system, by developing web-based interface that allow user to conveniently input data, visualize the forecast and manage the system.

With this architectural design, we are poised to develop a powerful and user-friendly sales forecasting system that leverages customer behavior data for accurate predictions.



4.5 Sprint 1 Implementation

This phase will involve translating the design into functional code and developing the initial features.

- **Data Collecting and Processing:** as part of the requirements gathering phase, we already obtained the dataset and performed the necessary operations to clean the dataset.
- **Variables Selection:** we did exploration on the dataset to find variables that have effects on the forecast such as WeaklySales, and we plan to explore in depth in this phase.
- **Algorithm Selection:** many algorithms is suitable for forecasting but we need to select the most appropriate for our system which match the busines goals.
- Testing: evaluating the accuracy of each model will help as choosing the appropriate algorithm, after that we need to adjust the model as needed to improve the accuracy.
- **Generating the Sales Forecast:** using the forecasting models to generate sales prediction for the future, by inputting relevant variables and utilizing the trained models, we will generate forecasts for the desired time periods.

- **Models we plan to test:**

  XGBoost, Artificial Neural Networks (ANN), and Genetic Programming. XGBoost is a powerful gradient boosting algorithm known for its accuracy and efficiency. ANN is a deep learning model capable of capturing complex patterns in the data. Genetic Programming is an evolutionary algorithm that evolves mathematical expressions to find the best fit for the sales forecasting problem.

4.6 Sprint 1 Testing and Evaluation

We plan to use several techniques to measure and evaluate the performance of the model:

4.6.1 Historical Data Comparison:

We will compare the projected results generated by the model with the actual outcomes from historical data and calculate the accuracy rate between them.

4.6.2 Cross-Validation:

Cross-validation is a technique used to assess the performance of a predictive model by dividing the data into subsets, training the model on one subset, and evaluating it on the remaining subset. It helps evaluate the model's accuracy and identify any overfitting or underfitting issues [12].

4.6.3 RMSE/MSE/MAE:

RMSE/MSE/MAE are a set of methods and techniques used to calculate the margin of error in predictions. They help quantify the level of accuracy or deviation between predicted values and actual values. These metrics provide valuable insights into the quality and reliability of the predictions made by a model [13].

n: number of data
e: error in function
t: counter

Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \vert e_i$

4.7 Conclusion

The Sprint 1 Design and Development phase laid the groundwork for our sales forecasting system, encompassing various design aspects and implementation steps. The use case diagrams provided a clear representation of the interactions and functionalities involving actors such as the program developer, retail companies, and the system itself. This helped identify the key roles and collaboration needed for accurate sales forecasting.

The aspects of design covered different components, including system engineering, data storage, integration, and dynamic aspects. These designs ensured a well-structured system architecture, efficient data storage and retrieval, seamless integration with other systems, and effective data preparation, model training, and sales forecasting processes.

The architectural aspect of the design outlined the three main components of the system: the data layer, model layer, and user interface layer. This design, supported by cloud-based platforms, ensures scalability, flexibility, and user-friendliness.

In terms of implementation, we focused on data collection and processing, variable selection, algorithm selection, testing, and generating sales forecasts. By performing these tasks, we obtained a cleaned dataset, identified relevant variables, selected suitable forecasting algorithms, and developed the initial features of the system.

Testing and evaluation techniques, such as historical data comparison, cross-validation, and RMSE/MSE/MAE metrics, will be employed to measure and assess the accuracy and reliability of the forecasting models.

Overall, the Sprint 1 design and development phase established a solid foundation for our sales forecasting system. It defined the system's functionalities, designed its architecture and components, implemented crucial features, and set the stage for further enhancements and iterations in subsequent sprints.

5. Chapter 5 Sprint 2 Development Report

In this chapter, we explore the intricate details of designing and developing an advanced sales forecasting system. This phase is crucial, establishing the foundation for implementing sophisticated artificial intelligence techniques.

The design phase is a pivotal segment, encompassing the conceptualization of the system architecture and the establishment of foundational aspects. Use-case diagrams, static and architectural aspect designs act as blueprints, delineating system interactions, and structural components. The use-case diagram is refined to encompass intermediate functional requirements, transitioning from abstract conceptualization to a tangible representation of system interactions.

The static aspect design showcases core classes encapsulating data-related aspects, data preprocessing, and model operations. The architectural aspect design organizes layers within the system governing data flow and processing. The dynamic aspect design reveals the internal dynamics governing the sales forecasting process, serving as a roadmap from raw data manipulation to the final forecasting.

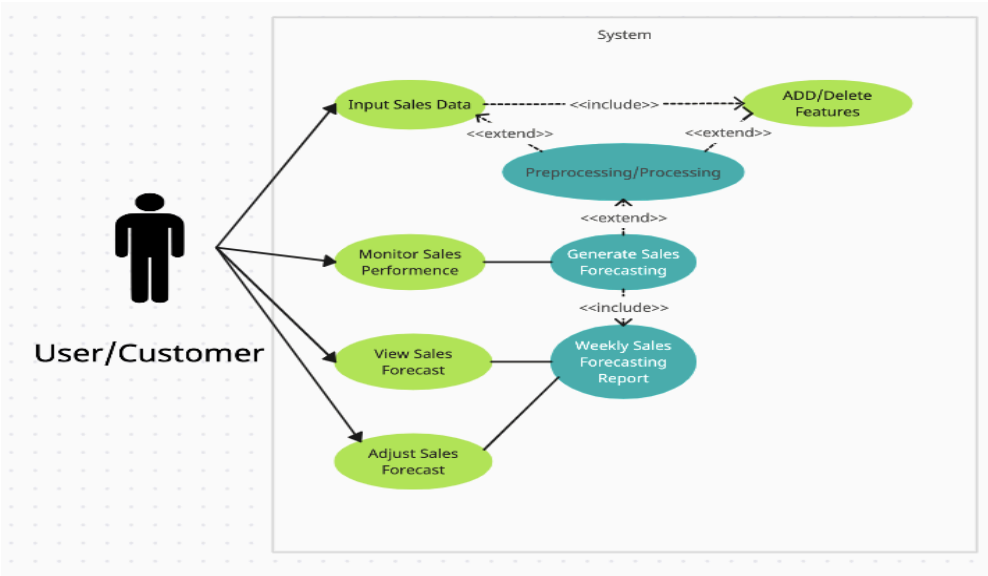
Moving from design to development, Sprint 2 is dissected to uncover implementation intricacies. This journey encompasses data preprocessing, feature engineering, and the exploration of modeling paradigms such as Linear Regression and Long Short-Term Memory (LSTM). Each coding and implementation step is elaborated, offering insights into decisions made to enhance system adaptability and intelligence.

The parts that follow present results of thorough testing and evaluation of the implementation. In-depth analysis of LSTM and linear regression models reveals their strengths and potential improvements. These stages serve as checkpoints in the development process and serve as accelerators for ongoing progress.

The construction of an intelligent sales forecasting system is demonstrated in Chapter 5 by the careful planning, strategic design, and methodical development that went into it. The narrative captures the use of technological prowess and intellectual rigor to present a system positioned at the nexus of innovation and functionality.

5.1 Sprint 2 Use-Case Diagram

During this phase of development, we have meticulously refined our use-case diagram to align with the INTERMEDIATE functional requirements. The revised diagram serves as an illustrative representation of the complex interactions within the system, highlighting four key actor's integrals to its functionality.



1. **Input Sales Data:**

    - **Initiating the Process:** The Input Sales Data actor assumes a pivotal role in the commencement of the sales forecasting process. By presenting raw sales data, this actor serves as the instigator for a sequence of operations leading to sophisticated forecasts. Beyond the mere submission of data, this actor possesses the nuanced capacity to introduce or exclude features, facilitating a flexible input process. This dynamic capability enables the system to adapt to diverse data scenarios.

    - **Extension to Preprocessing and Processing:** The actor's significance extends beyond data submission. The inclusion of feature addition or removal triggers subsequent stages in the process. It seamlessly integrates with the preprocessing phase, where the submitted data undergoes necessary transformations to enhance its suitability for analysis. Subsequently, the processed data enters the core processing stage, where advanced algorithms and models are applied for accurate sales forecasting.

2. **Monitor Sales Forecasting:**

    - **Real-Time Progress Monitoring**: The Monitor Sales Forecasting actor assumes a central role in providing real-time oversight of the sales forecasting process. This involves continuous surveillance of intricate operations within the system. By designating a dedicated actor for monitoring, users can stay abreast of progress, identify potential bottlenecks, and ensure the seamless flow of the forecasting pipeline.

- **Proactive Intervention:** The actor's involvement transcends passive observation. In the presence of anomalies or unexpected developments, the Monitor Sales Forecasting actor possesses the capability to proactively intervene. This ensures that any issues are promptly addressed, contributing to the overall efficiency and reliability of the forecasting system.

3. **View Sales Forecast:**

    - **Accessible Insights**: The View Sales Forecast actor provides end-users with a direct avenue to access the generated sales forecasts. This interaction is tailored for simplicity and user-friendliness, allowing stakeholders to effortlessly gain valuable insights. Through intuitive interfaces, users can navigate and explore forecasted data, empowering them to make informed decisions based on the system's predictions.

    - **Customized Presentation**: The user's ability to view sales forecasts is not a uniform experience. This actor ensures that the presentation of forecasted information is customizable, catering to the specific needs and preferences of individual users. Whether through visual representations, detailed reports, or specific time frames, the View Sales Forecast actor facilitates a tailored and sophisticated experience.

4. **Adjust Sales Forecast:**

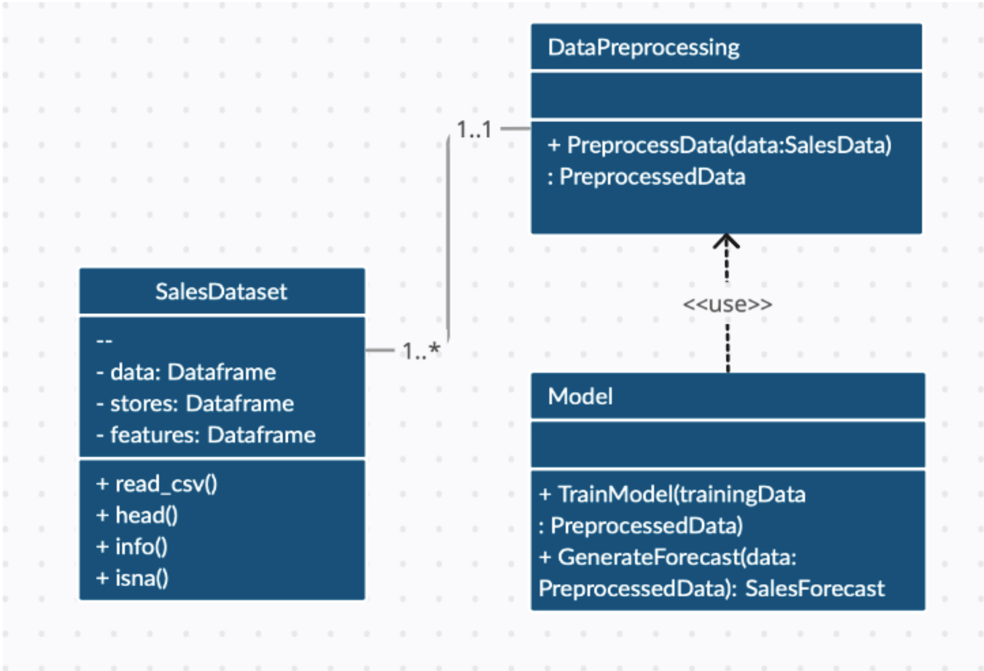
    - **User-Centric Adaptability:** The Adjust Sales Forecast actor embodies the user's authority to adapt and refine generated sales forecasts. This dynamic interaction acknowledges the inherent uncertainties and complexities in real-world scenarios. Users may possess domain-specific insights or external knowledge that can enhance the accuracy of forecasts. The actor provides a user-friendly interface for adjusting, fostering collaboration between the system and the user.

    - **Iterative Refinement:** The adjustability feature is not a one-time action; rather, it encourages an iterative refinement process. Users can interact with the system, assess the impact of adjustments, and iteratively refine forecasts. This feedback loop enhances the adaptability of the forecasting system, rendering it a valuable tool for dynamic business environments.

These interactions play a vital role in the system's operational framework, contributing to the accuracy and adaptability of the sales forecasting process. Notably, the Weekly Sales Forecasting Report is included in this diagram, reflecting the outcome.

5.2 Sprint 2 Aspect Design

In the context of Sprint 2 Aspect Design, our system's structural framework is meticulously portrayed through an exhaustive class diagram. This visual representation serves as a formal blueprint of the static aspect design, highlighting three pivotal classes that collectively constitute the foundational structure of our system.



1. **SalesDataset Class:**

    - **Data-Oriented Functionality:** The SalesDataset Class assumes a central role in overseeing the data-related facets of our system. This class functions as a repository for vital attributes, encompassing data, stores, and features. Operational functionalities include the proficient handling of tasks such as reading CSV data, presenting data through visualization mechanisms, and executing meticulous information checks to uphold data integrity.

    - **Interconnected Relationships:** Within the structural framework, the SalesDataset Class establishes a one-to-one relationship with the DataPreprocessing class. This relationship signifies a seamless flow of information, ensuring that the preprocessing stage aligns cohesively with the data-centric functionalities encapsulated by the SalesDataset Class.

2. **DataPreprocessing Class:**

   - **Data Transformation Proficiency**: The DataPreprocessing Class serves as a key player in data preprocessing within our system. Its core operation, PreprocessData, represents a sophisticated orchestration of tasks aimed at refining raw data into a state suitable for advanced analysis. This class, through its one-to-many relationship with the SalesDataset Class, exhibits a capacity to handle diverse datasets with efficiency and precision.

   - **Collaborative Integration:** The one-to-many relationship underscores a collaborative integration between the SalesDataset and DataPreprocessing classes. It signifies the adaptability of the system to handle multiple datasets, allowing for a flexible and dynamic preprocessing process tailored to specific data characteristics.

3. **Model Class:**

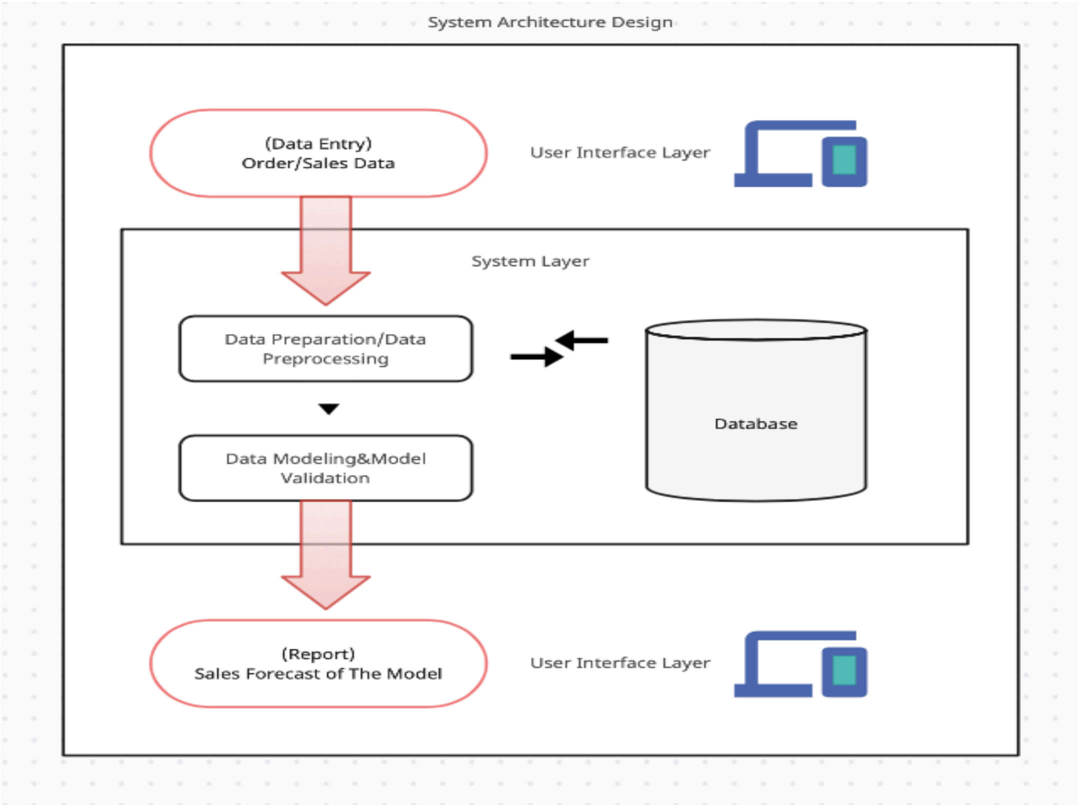
   - **Model-Driven Operations:** The Model Class serves as the linchpin for model-related operations within our system. Its multifaceted responsibilities encompass the training of models and the generation of forecasts. Leveraging the capabilities of the DataPreprocessing class, this class orchestrates a seamless transition from preprocessing raw data to facilitating model-driven insights.

   - **Interdependence with DataPreprocessing:** The strategic interdependence between the Model and DataPreprocessing classes denotes a cohesive synergy in the system's structural design. This symbiotic relationship ensures that data preprocessing, a critical precursor to modeling, transpires seamlessly, fostering an environment conducive to the generation of accurate and reliable forecasts.

In conclusion, the Sprint 2 Aspect Design unveils a meticulously crafted structural framework for our system, embodied by an exhaustive class diagram. This visual representation serves as a formal blueprint, elucidating the static aspect design with a focus on three pivotal classes that collectively underpin the foundational structure of our system.

5.3 Sprint 2 Architectural Aspect Design

Our architectural design is well-organized. As viewed in the diagram, this structured setup effectively coordinates data flow inside the system.



1. **Data Entry (User Interface Layer):**

    - **User Interaction Hub:** The Data Entry layer, situated at the User Interface layer, is instrumental in providing users with a dedicated space to input sales data into the system. This layer functions as the primary interface through which users interact with the system, ensuring a seamless and user-friendly experience in data submission.

2. **Data Preparation/Data Preprocessing (System Layer):**

    - **Preprocessing Nexus:** The Data Preparation layer, entrenched within the System layer, assumes responsibility for the intricate domain of data preprocessing. This involves a series of systematic operations to refine raw data and prepare it for subsequent analyses. Furthermore, this layer boasts privileged access to the system's database, affording it the capability to manipulate data with precision and efficacy.

- **Strategic Database Integration:** The seamless integration with the database empowers the Data Preparation layer to execute data manipulations with strategic finesse. This layer ensures not only the integrity of the data but also its optimal format for downstream processes, thereby contributing to the overall efficiency of the system.

3. **Data Modeling & Model Validation (System Layer):**

   - **Modeling Profundity:** Nestled within the System layer, the Data Modeling & Model Validation is the epicenter for sophisticated data modeling and rigorous model validation processes. This layer plays a pivotal role in shaping the models that underpin our forecasting system, utilizing advanced algorithms to extract meaningful patterns and insights from the preprocessed data.

   - **Validation Rigor:** The system strongly emphasizes model validation in conjunction with data modeling. To confirm the dependability and accuracy of the created models, stringent validation procedures are implemented. As a result, the projections generated by the system are guaranteed to be trustworthy and informative, reaching the highest requirements of analytical precision.

4. **Report (User Interface Layer):**

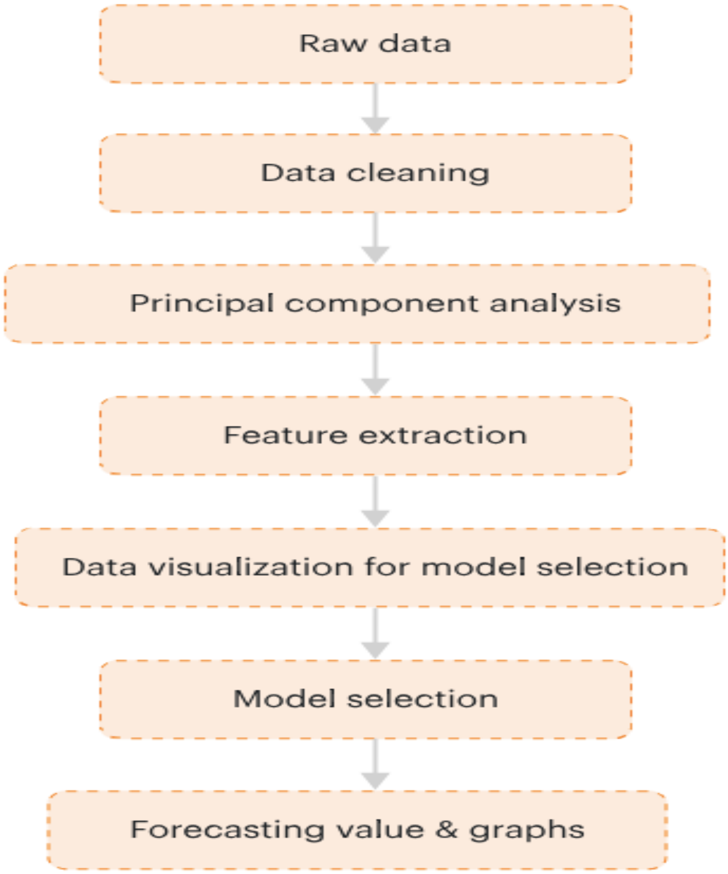
   - **Final Presentation Stage:** The Report layer, situated at the User Interface stratum, marks the culmination of the system's endeavors. It serves as the final presentation stage where the sales forecasts, meticulously generated through the preceding layers, are presented to end-users. This layer is dedicated to delivering insights in a comprehensible and visually appealing manner, catering to the diverse informational needs of stakeholders.

   - **User-Centric Reporting**: The design of the Report layer is inherently user-centric, acknowledging the diverse preferences of end-users in consuming forecasted information. Whether through graphical representations, detailed reports, or specific timeframes, this layer ensures a tailored and refined presentation of forecasts, enhancing the overall user experience.

In closing, this architectural design not only produces an effective composition that assures the continuous flow of data and the efficient operation of the system, but also clearly defines the distinct roles of each layer. It embodies a careful and defined system architectural approach, which is necessary for the achievement of our forecasting goals.

5.4 Sprint 2 Dynamic Aspect Design

In the Sprint 2 Dynamic Aspect Design, a meticulous elucidation of the internal dynamics of the system is undertaken to provide a comprehensive understanding of the sequential operations inherent in the process of generating sales forecasts.



1. **Raw Data Handling:**

    - **Inceptive Stage:** The dynamic process commences with the acquisition and ingestion of raw data, signifying the foundational phase of the forecasting endeavor. This initial stage involves the systematic retrieval and incorporation of pertinent sales data, laying the groundwork for subsequent analytical procedures.

2. **Data Cleaning Procedures:**

    - **Refinement Stage:** Following the raw data acquisition, a critical phase ensues wherein meticulous data cleaning procedures are employed. This involves the identification and rectification of anomalies, outliers, and missing values to ensure the integrity and quality of the dataset. The emphasis here is on fostering a dataset that is robust and conducive to accurate forecasting analyses.

3. **Principal Component Analysis (PCA):**

    - **Dimensionality Reduction:** The dynamic process advances to the application of Principal Component Analysis, a sophisticated statistical technique. PCA facilitates dimensionality reduction within the dataset, extracting essential features and patterns that contribute significantly to the variability within the data. This stage aims to enhance the efficiency of subsequent forecasting algorithms by focusing on the most influential components.
    -

4. **Feature Extraction Operations:**

    - **Information Abstraction**: Subsequent to PCA, the feature extraction stage unfolds, wherein the system discerns and abstracts critical features from the dataset. This process is pivotal in distilling relevant information that is instrumental in shaping the subsequent forecasting models. The goal is to capture the essence of the data in a streamlined manner.

5. **Data Visualization for Model Selection:**

    - **Strategic Representation**: A crucial juncture in the dynamic process involves data visualization tailored for model selection. This step encompasses the presentation of pertinent data patterns and relationships through graphical representations. The aim is to furnish stakeholders with a visually intuitive understanding, facilitating informed choices in the subsequent model selection phase.

6. **Model Selection Procedures:**

    - **Algorithmic Decision-Making:** The dynamic progression leads to the model selection phase, wherein algorithms suited to the specific forecasting objectives are judiciously chosen. This stage involves a thorough evaluation of algorithmic efficacy based on the preprocessed and visualized data, ensuring alignment with the intricacies of the dataset and the forecasting requirements.

7. **Forecasting Value & Graph Generation:**

    - **Culminating Stage:** The final dynamic stage culminates in the generation of forecasting values and graphical representations. Leveraging the selected forecasting model, the system extrapolates future sales values. Concurrently, graphical representations are crafted to visually communicate forecasted trends, providing stakeholders with comprehensive insights for strategic decision-making.

In conclusion, this detailed exposition of the Sprint 2 Dynamic Aspect Design underscores the intricate sequence of operations inherent in the sales forecasting process. Each stage, from raw data handling to forecasting value generation, contributes meaningfully to the overall efficacy of the system's forecasting capabilities.

5.5 Sprint 2 Implementation

In the second sprint of development, our primary objective was the implementation of two distinct approaches: Linear Regression and LSTM (Long Short-Term Memory). These methodologies were thoughtfully selected to explore diverse avenues for modeling sales data.

The implementation journey commenced with meticulous data preprocessing. The dataset, sourced from "data.csv," underwent thorough processing steps. Categorical variables, such as the "Type" and "IsHoliday" columns, were transformed into numerical representations. This process involved mapping categorical values and encoding Boolean values, ensuring compatibility with our modeling techniques.

Further, the temporal aspect of the data was addressed by splitting the "Date" column into separate components: year, month, and day. This transformation not only facilitates better understanding but also equips our models to capture temporal nuances in sales patterns.

Following preprocessing, the dataset was partitioned into training and testing sets using the widely adopted train-test split technique. The separation of features (X) and target variable (Y) was executed systematically, setting the stage for model training.

The implementation phase incorporated both Linear Regression and LSTM models. Linear Regression, a conventional yet powerful technique, allows us to understand linear relationships between input features and weekly sales. In contrast, LSTM, a specialized neural network architecture, excels in capturing temporal dependencies within the data, offering a more nuanced approach to modeling.

Despite the absence of specific numerical metrics at this stage, the initial results indicated that our models did not achieve the desired level of accuracy. However, this outcome is integral to the iterative development process, prompting thoughtful adjustments and refinements in subsequent iterations.

5.6 Sprint 2 Testing & Evaluation

The testing and evaluation phase involved a thorough analysis of the implemented models, specifically Linear Regression and LSTM. Performance metrics were assessed to gauge the efficacy of our models, providing valuable insights into their strengths and areas for improvement.

While the models did not deliver optimal results at this juncture, this is not a cause for discouragement. Instead, it serves as an anticipated aspect of the iterative development process, where challenges become steppingstones for refinement.

The significance of these preliminary findings lies in their contribution to the overall enhancement of our sales forecasting system. These insights, garnered from the testing and evaluation phase, will guide us in making informed decisions for subsequent model adjustments and enhancements.

As we advance through the development stages, our commitment to continuous improvement remains steadfast. The iterative nature of our approach ensures that challenges encountered in the testing phase become catalysts for the evolution and optimization of our sales forecasting system.

5.7 Conclusion

In conclusion, the pursuit of developing a robust and intelligent sales forecasting system has been a journey marked by strategic considerations, methodical implementations, and insightful evaluations. Through the iterative process, we have navigated the intricacies of employing diverse artificial intelligence methodologies, specifically Linear Regression and Long Short-Term Memory (LSTM), to model and predict weekly sales.

The foundational phases, spanning use-case definition, refined aspect designs, and architectural considerations, laid the groundwork for a comprehensive and structured system. The use-case diagram encapsulated the essential interactions, ensuring that our system is poised to respond adeptly to the intricacies of sales forecasting. The static aspect design, represented by core classes, fostered a structural foundation conducive to data management and model operations. Architectural and dynamic aspect designs delineated the layers of system functionality and the internal dynamics guiding the forecasting process, respectively.

Sprint 2, focusing on implementation, delved into the nuances of data preprocessing, feature engineering, and the deployment of both traditional and neural network models. Linear Regression, offering interpretability, and LSTM, capitalizing on temporal dependencies, were strategically chosen to explore the breadth of modeling possibilities.

Testing and evaluation, a pivotal phase, illuminated the strengths and areas of improvement for our models. While the desired level of accuracy was not immediately attained, the iterative ethos embedded in the development process transforms challenges into catalysts for refinement.

In the continuous pursuit of excellence, these preliminary insights lay the foundation for future iterations and advancements. The development of sophisticated artificial intelligence systems necessitates a steadfast commitment to refinement, learning, and adaptability. As we move forward, these experiences become integral components of our trajectory toward an intelligent, reliable, and adaptive sales forecasting system.

6. Chapter 6 Sprint 3 Design & Development

Sprint 3 is a transformative chapter in our Sales Forecasting System's evolution, witnessing strategic decisions, meticulous refinement, and innovative implementations. This report navigates through each facet, showcasing how they contribute to crafting a robust and responsive forecasting solution.

The use-case diagram underwent critical refinement, elevating abstraction and emphasizing direct user interactions. This enhanced clarity and conciseness, providing a comprehensive overview of the system's functionalities.

Aspect Design, depicted through a comprehensive class diagram, established a cohesive and modular architecture. DataManager, ForecastGenerator, and Dashboard classes collaboratively form the system's structural bedrock.

Architectural Aspect Design maintained a well-structured layout, ensuring continuous data flow and efficient system operation. No alterations were made, underscoring its ongoing utility and effectiveness.

Dynamic Aspect Design detailed the internal dynamics, from raw data processing to forecasting output generation, emphasizing the sustained usefulness of the established dynamic aspect design.

The Dashboard Interface empowers stakeholders with visual insights into sales trends. While the current set of visualizations provides a valuable starting point, there is room for enrichments as the project progresses.

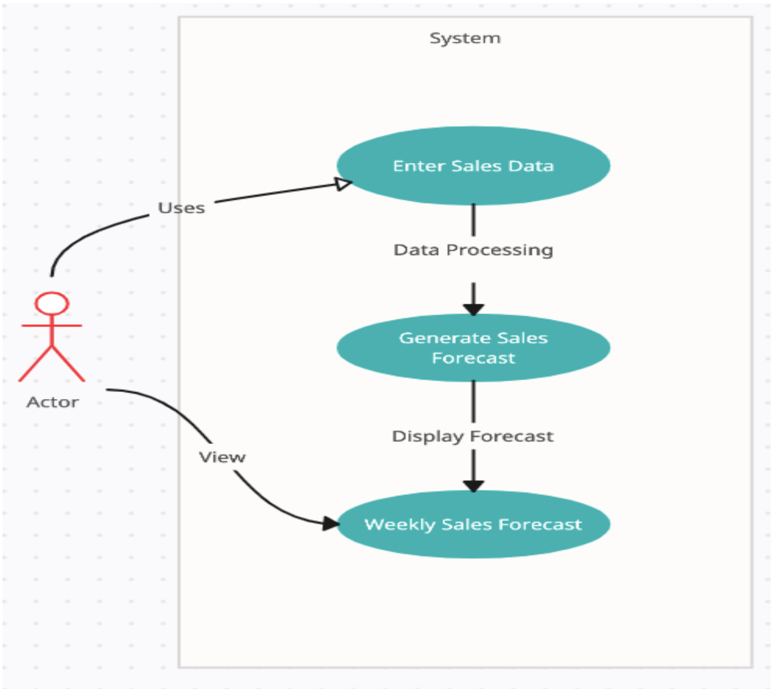
Implementation in Sprint 3 focused on meticulous refinement of four primary models: Linear Regression, Random Forest, K-Nearest Neighbors, and XGBoost. Feature engineering, anomaly handling, and external factor integration positioned our system for adaptability and accuracy.

Testing and Evaluation rigorously scrutinized the refined forecasting models, providing insights for future iterations. Challenges and opportunities surfaced, with a visual comparison plot aiding in the nuanced comprehension of model strengths and weaknesses.

As we conclude Sprint 3, our Sales Forecasting System stands as a refined and resilient solution, strategically poised for the challenges of forecasting in a rapidly evolving business landscape. The groundwork laid in this sprint positions our system as a dynamic and adaptable ally, reflecting a commitment to innovation, optimization, and a relentless pursuit of forecasting prowess.

6.1 Sprint 3 Use Case Diagram

In the concluding phase of our development process, specifically during the third sprint, we carried out additional refinements to perfect the use-case diagram for the Sales Forecasting System. Noteworthy is the strategic decision made in Sprint 2, where we opted to elevate the level of abstraction for the use case. This involved a concentrated focus on direct user interactions, ensuring a comprehensive and streamlined depiction of the final use case.



1. **Changes Made in Sprint 3:**

   Sprint 3 involved a critical revision of the use-case diagram to enhance its clarity and conciseness. We transitioned towards a more high-level representation, streamlining the focus to emphasize direct user interactions.

2. **Actors and System Boundary:**

   The primary actor, denoted as "Actor," symbolizes a generic user or stakeholder who interacts with the Sales Forecasting System.
   The "Sales Forecasting System" rectangle delineates the system boundary, encapsulating key use cases integral to the core functionality.

3. **Use Cases:**

   Enter Sales Data: The "Enter Sales Data" use case signifies the initial step where the Actor interacts with the system to input sales data. The arrow from Actor to Enter Sales Data signifies that the Actor employs this use case as part of their interaction with the system.

   Generate Sales Forecast: After the entry of sales data, the system engages in data processing to generate sales forecasts. The arrow from Enter Sales Data to Generate Sales Forecast signifies the logical sequence, highlighting the dependency on successful sales data entry.

   Weekly Sales Forecast (View): The "Weekly Sales Forecast" use case enables the Actor to view the generated sales forecasts. The arrow from Generate Sales Forecast to View Weekly Forecast signifies that the display of the weekly sales forecast is initiated after the successful generation of forecasts.

   Actor Interaction: The arrow from Actor to View Weekly Forecast underscores that the Actor, as the primary user, has the capability to initiate the action of viewing the weekly sales forecast.

4. **Logical Flow:**

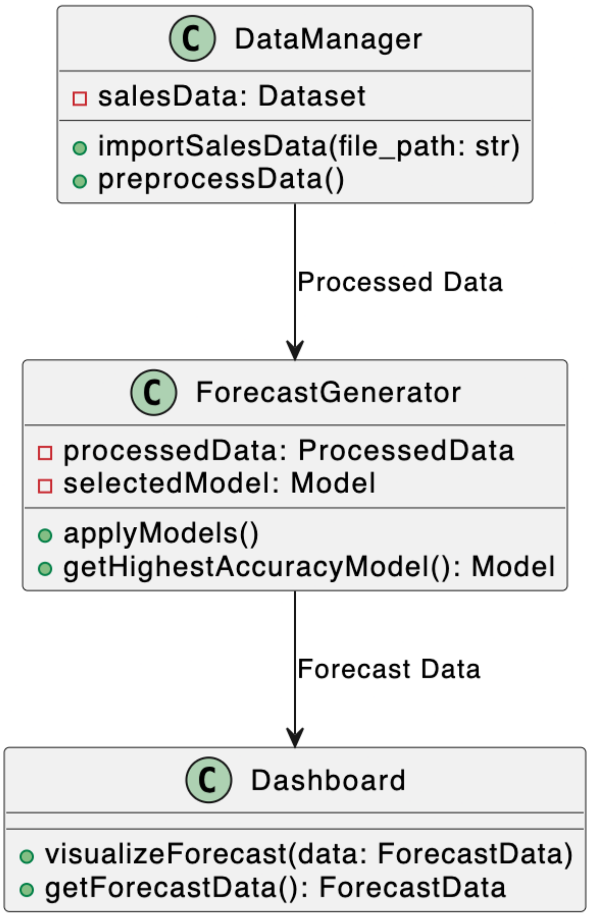
   The directional arrows serve the purpose of conveying the logical flow of actions, emphasizing the sequential nature of use cases. Specifically, the flow indicates that sales data entry precedes the generation of forecasts, and the viewing of the weekly forecast is contingent on successful data processing.

Sprint 3's use-case diagram builds upon the refinements made in Sprint 2, providing a comprehensive and detailed overview of the Sales Forecasting System's functionalities. This diagram serves as a valuable reference for understanding the logical flow and user interactions within the system, facilitating effective communication and development planning.

For a more detailed understanding of use case modeling and refinement, Cherfi, Akoka, and Comyn-Wattiau propose a quality-based approach that combines quality metrics with transformation rules, guiding software designers through a general framework. This innovative approach, detailed in their paper, sheds light on enhancing the quality of use case models and can be considered in the ongoing refinement of our use-case diagram [14].

6.2 Sprint 3 Aspect Design

In the culminating phase of Sprint 3 Aspect Design, the structural foundation of our system is intricately delineated through a comprehensive class diagram. This visual representation serves as a formal guide, outlining the static aspect design and placing emphasis on three pivotal classes that collectively constitute the fundamental framework of our system.



1. **DataManager Class:**

   - Data Management Operations: The DataManager class assumes a pivotal role in managing the data-centric operations of our Sales Forecasting System. It encapsulates essential attributes, including the salesData variable, representing the dataset. The class offers functionalities such as importing sales data from a file (importSalesData) and performing preprocessing on the imported data (preprocessData).

   - Data Flow Relationships: DataManager establishes a unidirectional relationship with the ForecastGenerator class. This relationship signifies the flow of processed data from DataManager to ForecastGenerator, ensuring a seamless transition of data for further forecasting operations.

2. **ForecastGenerator Class:**

   - Forecast Generation and Model Management: The ForecastGenerator class serves as the core engine for generating sales forecasts. It encapsulates attributes such as processedData representing the data after preprocessing and selectedModel representing the chosen machine learning model. Functionalities include applying multiple models (applyModels) and determining the model with the highest accuracy (getHighestAccuracyModel).

   - Inter-Class Collaboration: ForecastGenerator interacts with both DataManager and Dashboard classes. The relationship with DataManager ensures a streamlined flow of processed data, while the association with Dashboard facilitates the visualization of forecasted data.

3. **Dashboard Class:**

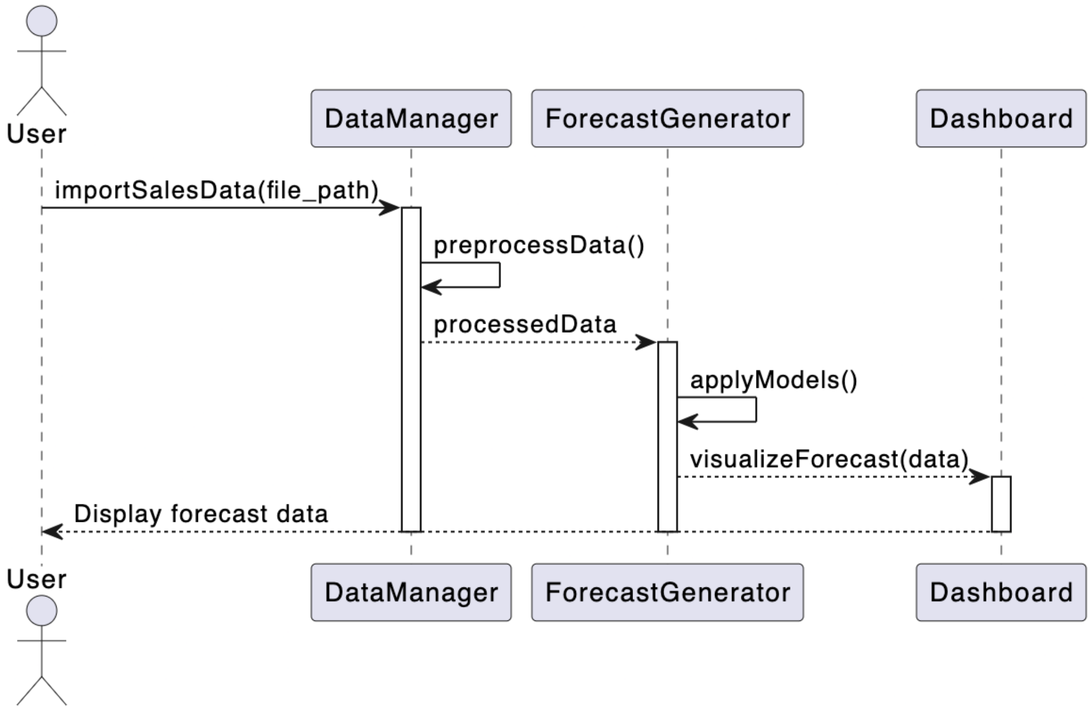
   - Visualization and Data Presentation: The Dashboard class specializes in visualizing forecasted sales data. It offers the visualizeForecast method to create visual representations of the forecast. Additionally, the getForecastData method enables external components to retrieve forecasted data for further analysis.

   - Collaborative Integration: The Dashboard class collaborates with ForecastGenerator to obtain forecast data. This collaboration ensures that the dashboard receives accurate and up-to-date information for visualization, fostering a dynamic and responsive user interface.

In conclusion, Sprint 3's class diagram delineates a comprehensive structural design for our Sales Forecasting System. The interplay between DataManager, ForecastGenerator, and Dashboard classes establishes a cohesive and modular architecture, facilitating efficient data management, forecast generation, and visualization functionalities. This visual representation serves as a crucial reference for implementing the dynamic aspects of our system, emphasizing collaboration and adaptability.

6.3 Sprint 3 Dynamic Aspect Design

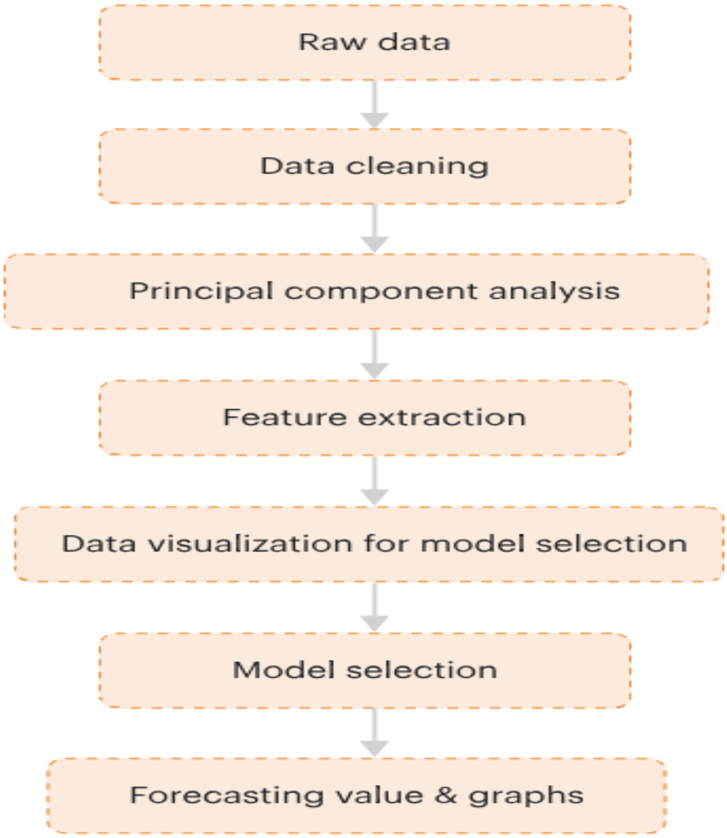
In the finalization of Sprint 3 Dynamic Aspect Design, a meticulous examination of the system's internal dynamics was conducted. The ensuing two diagrams succinctly depict the sequential operations governing the generation of sales forecasts.



1. **Actors:**
    - **User**: Represents an external entity interacting with the system.
2. **Participants:**
    - **DataManager**: Manages the import and preprocessing of sales data.
    - **ForecastGenerator**: Utilizes processed data to generate sales forecasts using various models.
    - **Dashboard**: Visualizes the generated forecasts and displays them to the user.

3. **Sequence of Actions:**
    - **Step 1 (User Interaction):**
        - The **User** initiates the process by importing sales data using the **importSalesData(file_path)** action.
        - This triggers the activation of the **DataManager**.
    - **Step 2 (Data Processing):**
        - The **DataManager** activates and performs data preprocessing via the **preprocessData()** action.
        - The processed data is then communicated to the **ForecastGenerator**.
    - **Step 3 (Forecast Generation):**
        - The **ForecastGenerator** activates and applies various models to generate forecasts using the **applyModels()** action.
        - Once the forecasting process is completed, the generated data is sent to the **Dashboard**.
    - **Step 4 (Data Visualization):**
        - The **Dashboard** is activated to visualize the forecast data using the **visualizeForecast(data)** action.
        - Finally, the visualized forecast data is displayed to the **User**.
4. **Deactivations:**
    - After the data is visualized, both the **Dashboard** and **ForecastGenerator** are deactivated, indicating the completion of the sequence of actions.

The diagram portrays a dynamic flow where the **User** triggers a chain of actions involving the **DataManager**, **ForecastGenerator**, and **Dashboard**. Each component performs its designated function in a well-orchestrated manner, showcasing the dynamic interactions within the Sales Forecasting System.



1. **Raw Data Processing:**

   - Initiating Phase: The dynamic process begins with the processing of raw data, marking the foundational phase of the forecasting endeavor. This initial stage involves the systematic processing and incorporation of pertinent sales data, setting the stage for subsequent analytical procedures.

2. **Data Cleansing Protocols:**

   - Refinement Stage: Following the processing of raw data, a critical phase ensues where meticulous data cleansing protocols are applied. This involves the identification and correction of anomalies, outliers, and missing values to ensure the integrity and quality of the dataset. The emphasis here is on fostering a dataset that is robust and conducive to accurate forecasting analyses.

3. **Principal Component Analysis (PCA):**

   - Dimensionality Reduction: The dynamic process advances to the application of Principal Component Analysis, a sophisticated statistical technique. PCA facilitates dimensionality reduction within the dataset, extracting essential features and patterns that significantly contribute to the variability within the data. This stage aims to enhance the efficiency of subsequent forecasting algorithms by focusing on the most influential components.

4. **Feature Extraction Procedures:**

   - Information Abstraction: After PCA, the feature extraction stage unfolds, wherein the system discerns and abstracts critical features from the dataset. This process is pivotal in distilling relevant information instrumental in shaping the subsequent forecasting models. The goal is to capture the essence of the data in a streamlined manner.

5. **Data Visualization for Model Selection:**

   - Strategic Representation: A crucial juncture in the dynamic process involves data visualization tailored for model selection. This step encompasses the presentation of pertinent data patterns and relationships through graphical representations. The aim is to furnish stakeholders with a visually intuitive understanding, facilitating informed choices in the subsequent model selection phase.

6. **Model Selection Protocols:**

   - Algorithmic Decision-Making: The dynamic progression leads to the model selection phase, wherein algorithms suited to the specific forecasting objectives are judiciously chosen. This stage involves a thorough evaluation of algorithmic efficacy based on the preprocessed and visualized data, ensuring alignment with the intricacies of the dataset and the forecasting requirements.

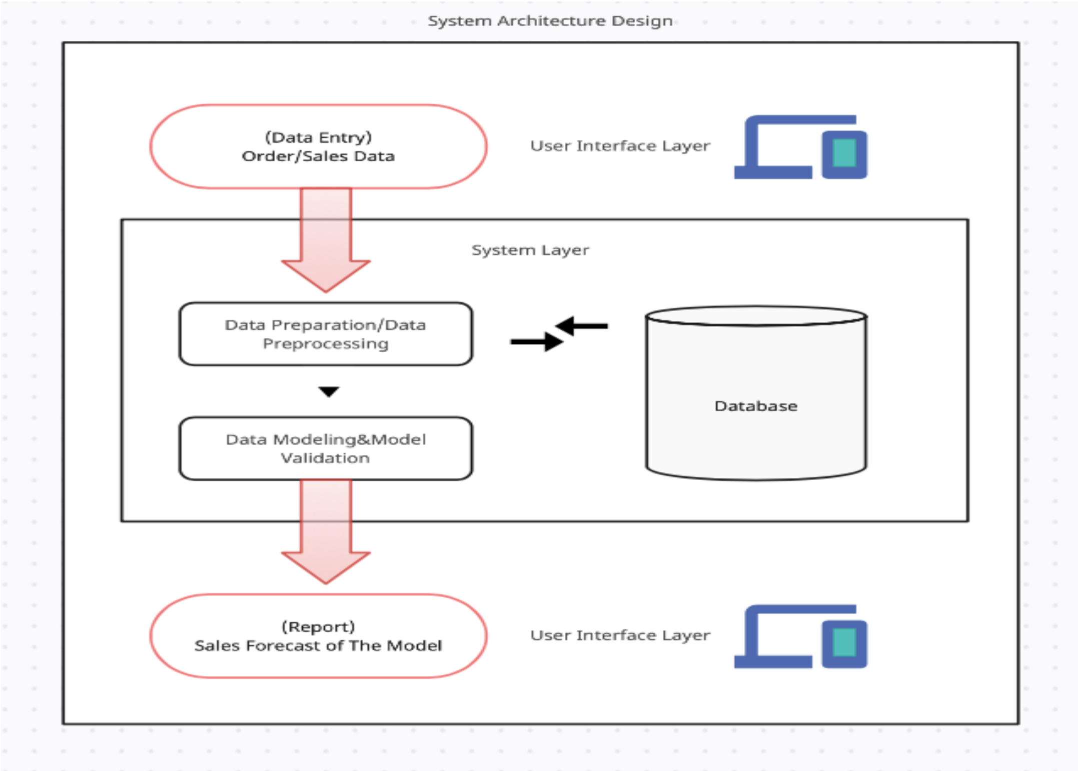
7. **Forecasting Output & Graph Generation:**

   - Culminating Stage: The final dynamic stage culminates in the generation of forecasting outputs and graphical representations. Leveraging the selected forecasting model, the system extrapolates future sales values. Concurrently, graphical representations are crafted to visually communicate forecasted trends, providing stakeholders with comprehensive insights for strategic decision-making.

In conclusion, our detailed exploration of Sprint 3's Dynamic Aspect Design highlights the intricate sequence of operations integral to the sales forecasting process. From raw data processing to forecasting output generation, each stage significantly contributes to the ongoing effectiveness of the system. In summary, the concise diagrams elucidate sequential operations, emphasizing the roles of DataManager, ForecastGenerator, Dashboard, and the User, marking a pivotal advancement in our system's development.

6.4 Sprint 3 Architectural Aspect Design

Our architectural layout stands resiliently well-structured. As manifested in the diagram, this organized arrangement adeptly manages the flow of data within the system. Notably, there were no modifications made in this sprint, affirming its ongoing effectiveness in fulfilling its intended purpose.



1. **User Input (User Interface Layer):**

    - User Engagement Hub: The User Input layer, positioned within the User Interface stratum, serves as a central hub for users to input sales data into the system. This layer acts as the primary interface, ensuring a smooth and user-friendly experience for data submission.

2. **Data Processing/Data Preprocessing (System Layer):**

    - Processing Hub: The Data Processing layer, entrenched in the System stratum, takes charge of the intricate domain of data preprocessing. Through systematic operations, it refines raw data for subsequent analyses. With privileged access to the system's database, this layer manipulates data with precision and efficacy.

    - Integrated Database Operations: Seamless integration with the database empowers the Data Processing layer to execute manipulations with strategic finesse. It upholds data integrity and ensures optimal formatting for downstream processes, enhancing the overall efficiency of the system.

3. **Data Modeling & Model Verification (System Layer):**

    - Modeling Excellence: Nestled within the System stratum, the Data Modeling & Model Verification layer serves as the epicenter for advanced data modeling and rigorous model validation processes. It shapes models using sophisticated algorithms to extract meaningful patterns from preprocessed data.

    - Validation Precision: The system places strong emphasis on model validation alongside data modeling. Stringent procedures confirm the dependability and accuracy of models, guaranteeing trustworthy and informative projections that meet the highest standards of analytical precision.

4. **Reporting (User Interface Layer):**

    - Final Presentation Platform: The Reporting layer, positioned at the User Interface level, marks the culmination of system efforts. It serves as the ultimate presentation stage where meticulously generated sales forecasts are presented to end-users. This layer focuses on delivering insights in a comprehensible and visually appealing manner to cater to diverse informational needs.

    - User-Centric Presentation: The design of the Reporting layer is inherently user-centric, recognizing diverse user preferences. Whether through graphical representations, detailed reports, or specific timeframes, this layer ensures a tailored and refined presentation of forecasts, enhancing the overall user experience.

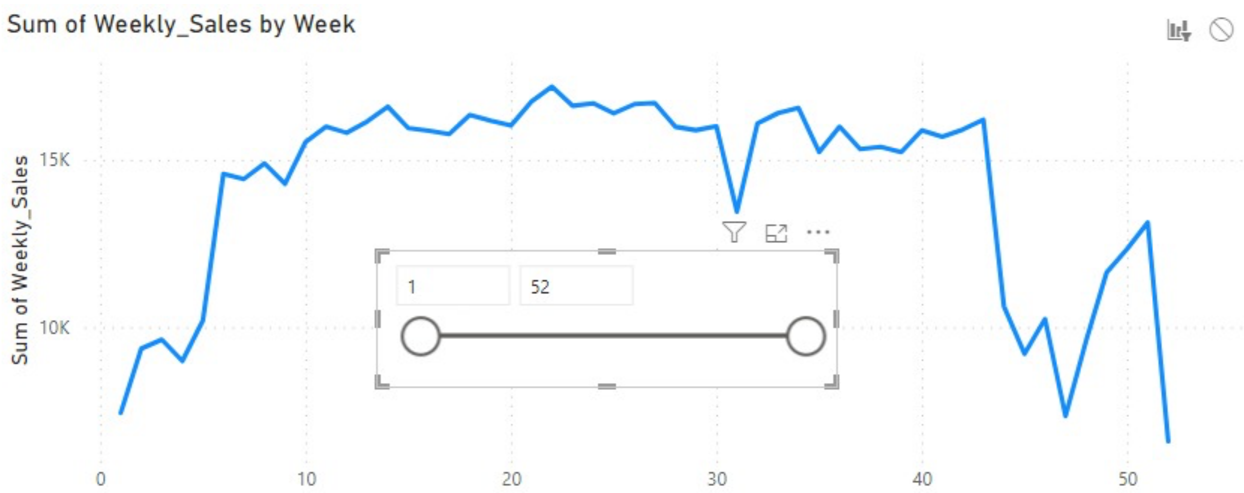
In summary, this architectural design maintains an effective structure, ensuring continuous data flow and efficient system operation. It distinctly defines the roles of each layer, embodying a thoughtful and defined architectural approach necessary for achieving our forecasting goals. No changes were made in this sprint, underscoring the ongoing utility and effectiveness of the established architectural design.

6.5 Sprint 3 Dashboard Interface

The Dashboard Interface of our Sales Forecasting System provides a user-friendly and intuitive platform for stakeholders to gain valuable insights into sales trends [15]. Leveraging visualization techniques, the interface presents a comprehensive view of the impact of various features on weekly sales, aiding informed decision-making. Here are the key visualizations available in the Dashboard:

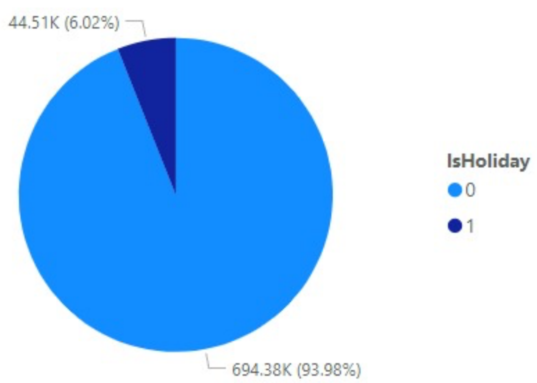
1. **Sum of Weekly Sales by Week:**

    - This plot depicts the overall trend of weekly sales over time. Users can easily observe fluctuations, identify peak periods, and understand the general trajectory of sales.



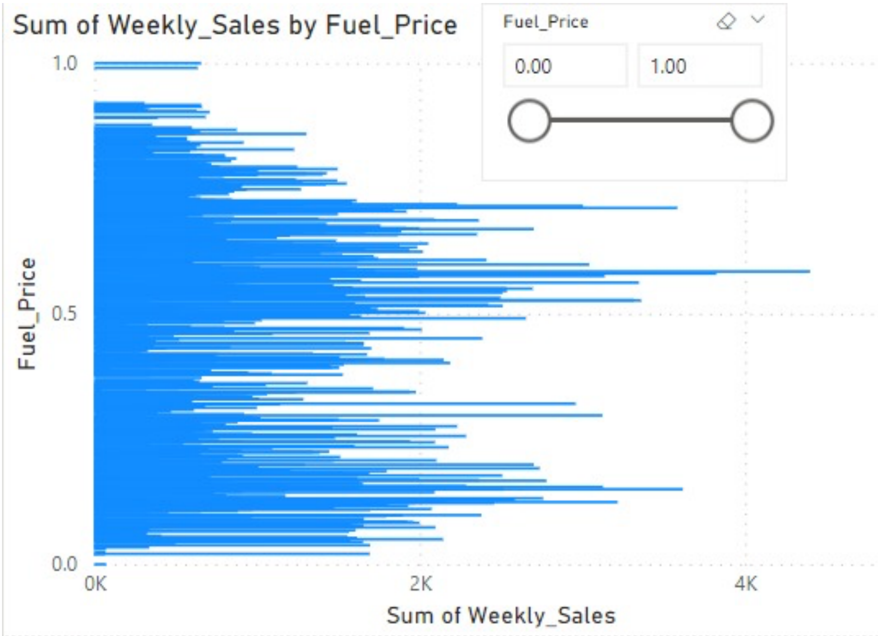
2. **Sum of Weekly Sales by IsHoliday:**

    - IsHoliday plays a significant role in sales dynamics. This visualization highlights the impact of holidays on weekly sales, providing insights into the correlation between holiday periods and increased sales.



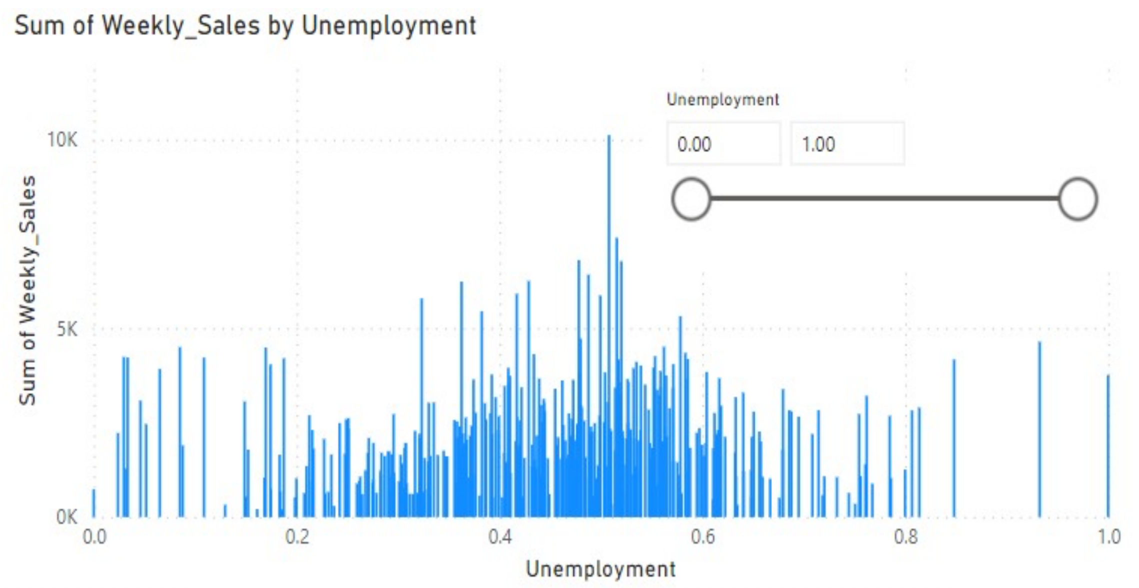
3. **Sum of Weekly Sales by Fuel Price:**

- Fuel prices can influence consumer behavior. This plot illustrates how changes in fuel prices correlate with fluctuations in weekly sales, helping users discern patterns and potential causative factors.



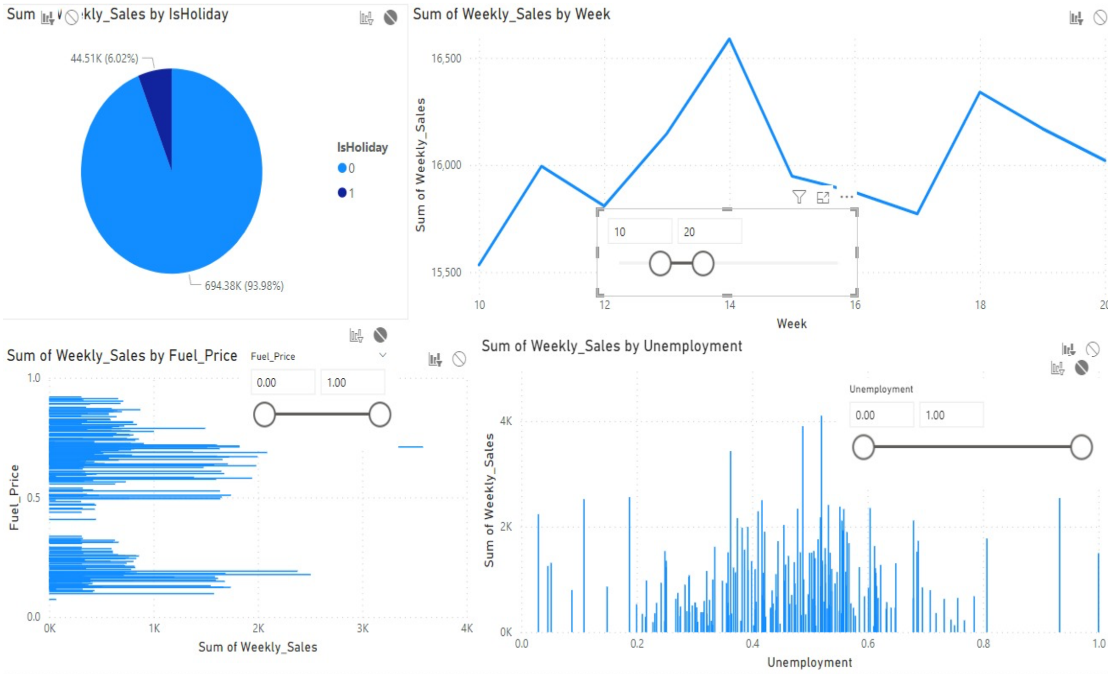
4. **Sum of Weekly Sales by Unemployment:**

- Unemployment rates can also influence sales trends. This visualization enables users to explore the relationship between unemployment levels and weekly sales, identifying patterns that may impact forecasting.



5. Custom Range of Weeks:

- Users can select a specific range of weeks to view a consolidated set of plots encompassing all relevant features. This dynamic feature allows for a more detailed exploration of how multiple factors interplay and contribute to sales variations.



These visualizations collectively empower users to make data-driven decisions by offering a preliminary insight into the underlying patterns and correlations present in the dataset. The Dashboard Interface serves as a central hub for initial explorations, paving the way for a more comprehensive and strategic approach to sales forecasting. While the current set of plots provides a valuable starting point, there is ample room for additional visualizations to enhance the depth and breadth of insights available to users. As the project progresses, the Dashboard Interface can be further enriched with additional plots and features, ensuring a continually evolving and robust platform for informed decision-making in the realm of sales forecasting.

6.6 Implementation

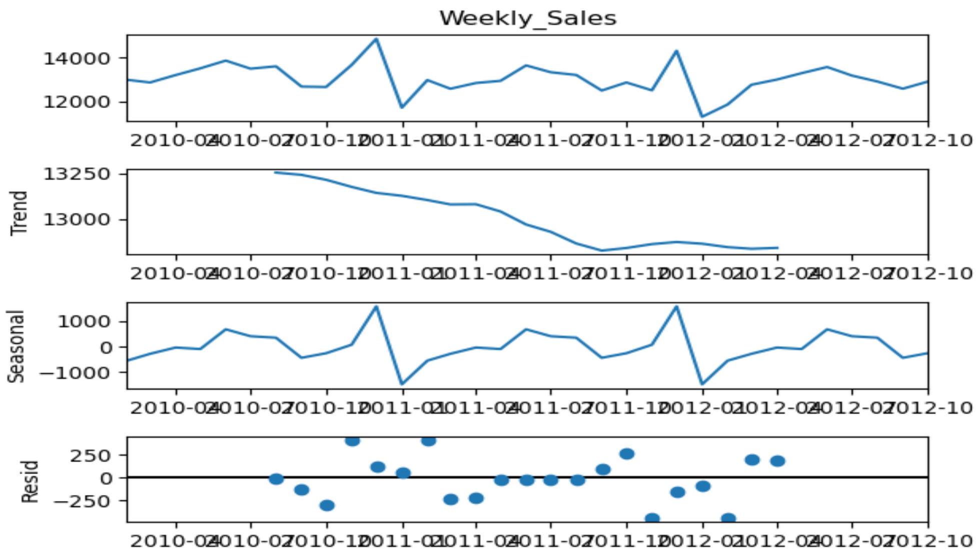
In the third sprint, our developmental focus shifted towards a meticulous refinement of our sales forecasting system. Drawing upon insights derived from the iterative evolution witnessed in Sprint 2, our endeavors concentrated on further optimizing the predictive capabilities of four primary models: Linear Regression, Random Forest, K-Nearest Neighbors, and XGBoost [16][17].

- **Model Refinement and Feature Engineering**

  A committed pursuit of precision led us to undertake a comprehensive model refinement process. Each model underwent meticulous fine-tuning, guided by nuanced insights gleaned during the rigorous testing and evaluation phase in Sprint 2. Our scope expanded to encompass advanced feature engineering techniques, aimed at extracting deeper insights from the dataset.
  Delving into the temporal dimensions initially addressed by the fundamental "Date" column split, our objective was to unveil more intricate time-related features. This strategic initiative sought to uncover subtle patterns and dependencies within the sales data, significantly enhancing the predictive capabilities of our models.

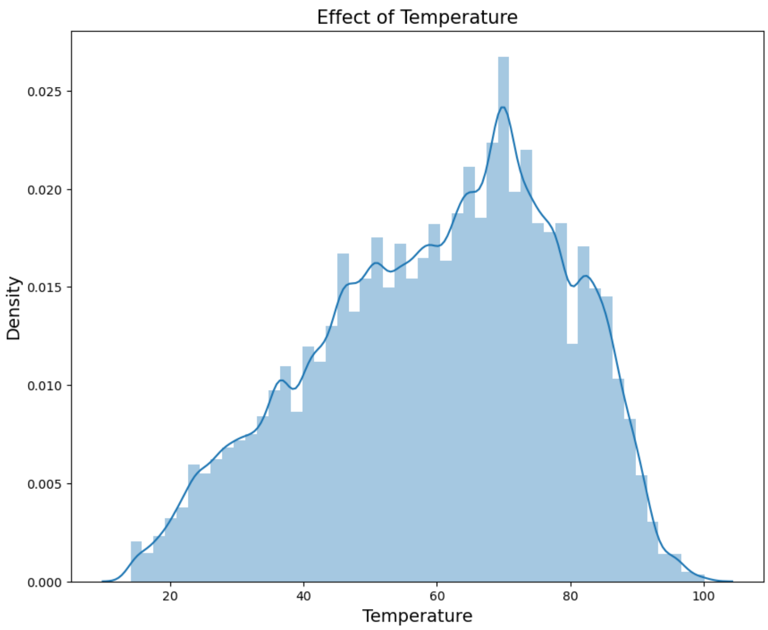
- **Handling Anomalies and Outliers**

  Recognition of the pivotal role played by anomalies and outliers in influencing forecasting accuracy prompted the implementation of robust techniques to detect and address these elements within the dataset. The overarching goal was to fortify our models against irregularities in the input data, fostering resilience and consistency in forecasts.



- **Integration of External Factors**

  Acknowledging the real-world impact of external factors on sales patterns, our data horizon expanded beyond the initial dataset. We seamlessly integrated additional external data sources into our forecasting models, strategically aiming to enrich predictive capabilities by considering variables extending beyond the confines of the original dataset.



- **Fine-Tuning for Robust Predictions**

  Our commitment to precision extended to fine-tuning every aspect of our models. Parameters were scrutinized, algorithms optimized, and ensemble methods explored to orchestrate a refined of predictions. This phase leveraged both domain expertise and tangible insights derived from the evaluation metrics of preceding sprints.

  Algorithmic model tuning, as illustrated in the study by Tuppi et al. has proven its reliability in semi-realistic scenarios. The positive outcomes observed in these controlled cases instill confidence in the capability of algorithmic models to yield robust results. As we delve into the complexities of semi-realistic cases, the success of model tuning becomes a beacon, guiding our approach and providing a strong foundation for transitioning into fully realistic scenarios [18].

- **Comprehensive Testing Iterations**

  As each refinement unfolded, comprehensive testing iterations were paramount. The models underwent rigorous testing against diverse scenarios and datasets, ensuring that the refined versions withstood the challenges posed by real-world variations. This iterative testing approach, integral to our development ethos, served as a litmus test for the adaptability and robustness of our forecasting system.

- **Strategic Documentation for Future Iterations**

  Documenting our refined models and the intricate steps taken during this sprint was a strategic imperative. The insights gained, decisions made, and challenges overcome were meticulously recorded. This documentation serves not only as a knowledge repository for our team but also as a guide for future iterations, embodying our commitment to continuous improvement.

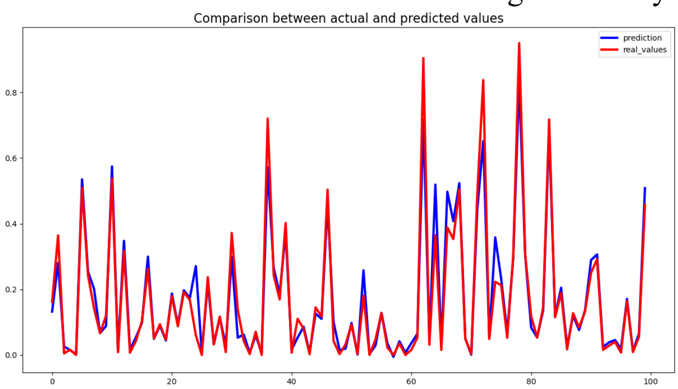
In this sprint, our journey towards a precision-driven sales forecasting system took center stage. The intricate dance of model refinement, feature engineering, anomaly handling, and external factor integration has positioned our system at the forefront of adaptability and accuracy. As we transition to the next phase, the refined models stand as a testament to our unwavering dedication to excellence in the field of sales forecasting.

6.7 Sprint 3 Testing and Evaluation

In the third sprint of our development journey, rigorous testing and evaluation constituted a pivotal phase, characterized by an in-depth scrutiny of our refined forecasting models. Leveraging the iterative ethos embedded in our approach, we meticulously assessed the performance of four distinct models: Linear Regression, Random Forest, K-Nearest Neighbors, and XGBoost. This thorough evaluation aimed to not only validate the efficacy of our model refinements but also provide valuable insights for future iterations.

- **Model-Specific Performance Analysis**
  1. *Linear Regression:*

     Linear Regression model, a conventional yet robust technique. The model exhibited commendable accuracy, achieving an R-squared value of 92.28%. The Mean Absolute Error (MAE) of 0.0301 and Root Mean Squared Error (RMSE) of 0.0590 underscored the model's proficiency in capturing weekly sales patterns. An examination of the model's predictions against actual values revealed a nuanced understanding of sales dynamics.

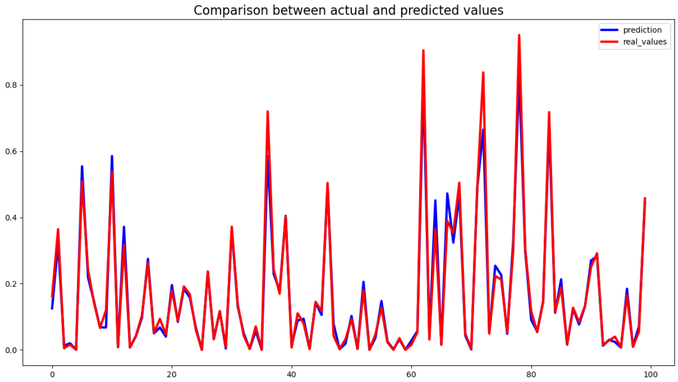



```
MAE 0.03007921529353659
MSE 0.0034837903884710327
RMSE 0.05902364262285947
R2 0.9228379352983277
```

Date	Actual	Predicted
2011-08-05	0.161661	0.132118
2010-07-09	0.364278	0.280514
2011-07-01	0.005003	0.025738
2012-01-06	0.015856	0.015174
2011-08-26	0.000318	0.001827

2. *Random Forest:*

Moving to the Random Forest model, we observed a notable improvement in accuracy, with an R-squared value of 97.91%. The model's precision in predicting weekly sales was evident in its low MAE of 0.0155 and RMSE of 0.0308. The model not only exhibited robustness in handling complex patterns but also showcased superior performance compared to the Linear Regression counterpart. The DataFrame visualization reinforced the model's capability to capture subtle variations in sales data.

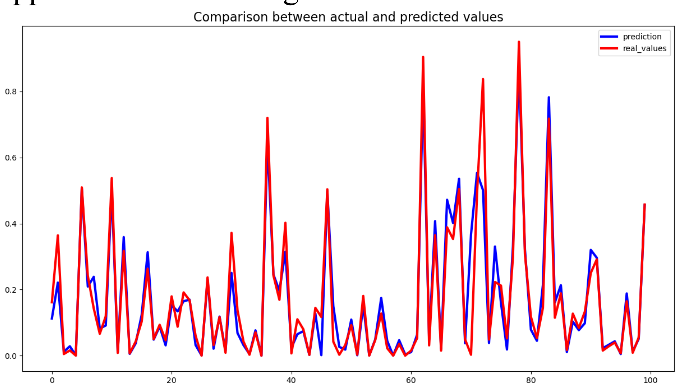



```
MAE 0.015479306651513208
MSE 0.0009456715119559565
RMSE 0.030751772501043845
R2 0.9790547984778172
```

Date	Actual	Predicted
2011-08-05	0.161661	0.125436
2010-07-09	0.364278	0.321535
2011-07-01	0.005003	0.012978
2012-01-06	0.015856	0.019651
2011-08-26	0.000318	0.000573

3. *K-Nearest Neighbors (KNN):*

Our evaluation extended to the KNN model, revealing an accuracy of 92.14%. While slightly trailing the Random Forest model, KNN demonstrated robustness, evident in its MAE of 0.0332 and RMSE of 0.0596. The model's predictions, as illustrated in the DataFrame, showcased a commendable alignment with actual sales values, albeit with a nuanced approach to forecasting.

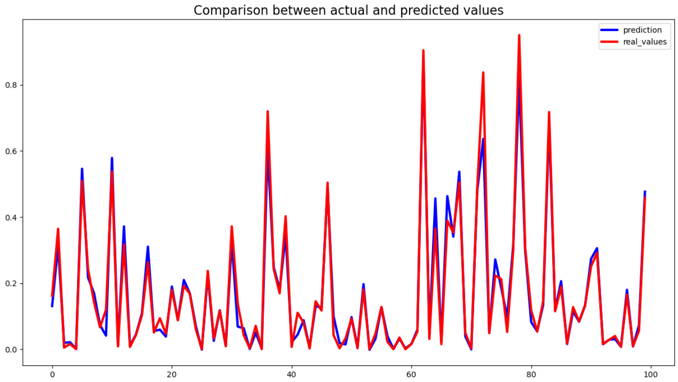



```
MAE 0.033150338631610654
MSE 0.003547108134077407
RMSE 0.059557603495082025
R2 0.9215859611622021
```

Date	Actual	Predicted
2011-08-05	0.161661	0.112559
2010-07-09	0.364278	0.221307
2011-07-01	0.005003	0.011921
2012-01-06	0.015856	0.028551
2011-08-26	0.000318	0.001063

4. *XGBoost:*

The evaluation culminated with the XGBoost model, which exhibited an impressive R-squared value of 97.35%. The model's nuanced understanding of temporal dependencies and intricate patterns was reflected in its low MAE of 0.0197 and RMSE of 0.0346. XGBoost demonstrated a holistic comprehension of sales dynamics, surpassing the accuracy of both Linear Regression and KNN. The DataFrame visualization further underscored the model's ability to generate precise sales forecasts.




```
MAE 0.01974456712971368
MSE 0.0011981157286570685
RMSE 0.03461380835240567
R2 0.9734634143804428
```

Date	Actual	Predicted
2011-08-05	0.161661	0.130563
2010-07-09	0.364278	0.313040
2011-07-01	0.005003	0.019386
2012-01-06	0.015856	0.021007
2011-08-26	0.000318	0.000722

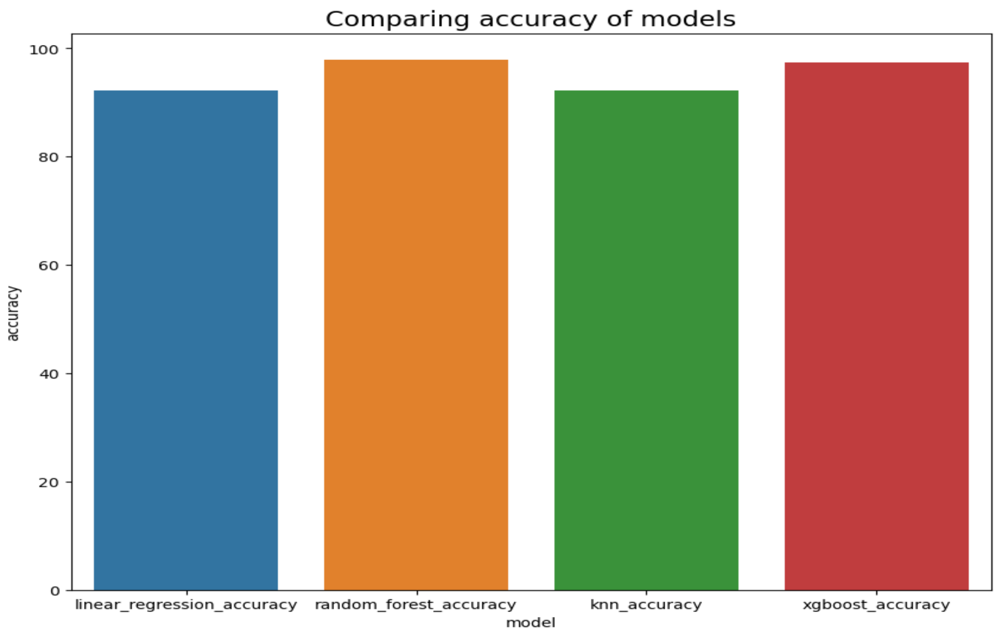
- **Challenges and Opportunities for Future Enhancements**

While our models exhibited commendable accuracy, the testing and evaluation phase unveiled challenges and opportunities for further enhancements. Notably, the impact of outliers and anomalies on forecasting accuracy requires ongoing attention. Future iterations may explore advanced anomaly detection techniques, ensuring the robustness of models in diverse scenarios.

Furthermore, the integration of external factors into our models showcased potential improvements. Continuous refinement of the selection criteria for external data sources and exploration of additional variables could further enrich the predictive capabilities of our forecasting system.

- **Models Comparison Section**

  The visual comparison plot below offers a graphical depiction of accuracy metrics, specifically R-squared values, across all models. This in-depth analysis aids in developing a nuanced comprehension of the inherent strengths and weaknesses of each model, providing stakeholders with valuable insights into their adaptability and precision in forecasting sales patterns.



In conclusion, the Sprint 3 Testing and Evaluation phase not only validated the enhanced capabilities of our forecasting models but also provided valuable insights for future iterations. The academic rigor applied in scrutinizing each model's performance, coupled with the documentation of challenges and opportunities, positions our forecasting system for continuous refinement and optimization in the pursuit of excellence.

6.8 Conclusion

In conclusion, our journey to develop a sophisticated Sales Forecasting System has been characterized by strategic decision-making, meticulous implementations, and insightful evaluations. Throughout this iterative process, we navigated through the intricacies of refining a use-case diagram, aspect designs, and architectural considerations, ensuring a comprehensive and well-structured system.

The core classes, as depicted in the static aspect design, provided a solid foundation for data management and model operations, while architectural and dynamic aspect designs outlined the layers of system functionality and the internal dynamics guiding the forecast.

Sprint 3, with its focus on further refinement and optimization, allowed us to delve into the nuances of model implementation, feature engineering, and system architecture. The comprehensive class diagram in the aspect design facilitated a modular and cohesive architecture, enhancing the efficiency of our Sales Forecasting System.

The implementation phase strategically refined four primary models: Linear Regression, Random Forest, K-Nearest Neighbors, and XGBoost, ensuring adaptability and accuracy in forecasting. Rigorous testing and evaluation provided valuable insights into model performance, paving the way for future enhancements.

While challenges surfaced, the iterative ethos ingrained in our development process positions these challenges as catalysts for refinement. As we move forward, these experiences become integral components of our trajectory toward an intelligent, reliable, and adaptive sales forecasting system. The groundwork laid in Sprint 3 serves as a cornerstone for continuous improvement, ensuring that our system evolves in tandem with the dynamic landscape of sales forecasting.

General Conclusion

In wrapping up the development of our Sales Forecasting System, our journey mirrors the essence of our project's overarching goal—creating a tailored software solution for precise sales predictions in the retail sector. We embarked on this endeavor recognizing the pivotal role accurate sales forecasting plays in refining inventory management, production planning, and distribution processes for retailers.

Much like the significance of our project's primary objective, our developmental path unfolded through strategic decision-making and meticulous implementation. Our focus extended to key components such as use-case diagrams, aspect designs, and architectural considerations, all converging to refine the system's foundational structures.

The core strength of our system lies in its well-defined architecture, exemplified by a use-case diagram that underwent critical refinement to ensure clarity and conciseness. A comprehensive class diagram further highlights the cohesive and modular architecture of pivotal classes—DataManager, ForecastGenerator, and Dashboard. This structural foundation ensures a seamless data flow and efficient system operation, emphasizing our commitment to creating a resilient solution.

Sprint 3 marked a transformative phase where our attention shifted to the meticulous refinement of forecasting models. Linear Regression, Random Forest, K-Nearest Neighbors, and XGBoost were strategically fine-tuned to enhance adaptability and accuracy. Feature engineering, anomaly handling, and the integration of external factors positioned our system as a dynamic and adaptable ally in the ever-evolving business landscape.

Testing and evaluation provided invaluable insights into the performance of our refined models, revealing commendable accuracy, and offering a nuanced understanding of sales dynamics. Challenges surfaced, providing opportunities for future enhancements. The integration of external factors showcased potential improvements, emphasizing the continuous refinement needed for optimal predictive capabilities.

Our Dashboard Interface emerged as a user-friendly platform, providing stakeholders with visual insights into sales trends. Visualizations, including weekly sales trends over time, the impact of holidays on sales, and the correlation between external factors and weekly sales, offered a comprehensive view. The system's adaptability was further emphasized by allowing users to select a custom range of weeks for detailed exploration.

As we conclude this developmental phase, the groundwork laid in Sprint 3 positions our Sales Forecasting System as a refined and resilient solution, strategically poised for the challenges of forecasting in a rapidly evolving business landscape. The iterative ethos ingrained in our development process ensures that challenges are viewed as catalysts for refinement, paving the way for continuous improvement. Our commitment to innovation, optimization, and a relentless pursuit of forecasting prowess remains unwavering, and this report stands as a testament to our dedication to crafting an intelligent, reliable, and adaptive sales forecasting system.

References

[1] Curtis, A., Lundholm, R., & McVay, S. (2014). Forecasting Sales: A Model and Some Evidence from the Retail Industry. *Contemporary Accounting Research*, 31, 581-608. https://doi.org/10.2139/SSRN.486162.

[2] Ji-wei, X., Yaohua, W., Qian, W., Li, L., & Hongchun, H. (2007). A Hybrid Sales Forecasting Method Based on Stable Seasonal Pattern Models and BPNN. *2007 IEEE International Conference on Automation and Logistics*, 2868-2872. https://doi.org/10.1109/ICAL.2007.4339070.

[3] CNBC. (2018). Toys 'R' Us' Poor Holiday Sales Cast Doubts on Its Future. https://www.cnbc.com/2018/01/30/toys-r-us-poor-holiday-sales-cast-doubts-on-its-future.html

[4] Vyas, R., & As, R. (2022). Seasonal Sales Prediction and Visualization for Walmart Retail Chain Using Time Series and Regression Analysis: A Comparative Study. *2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, 1-6. https://doi.org/10.1109/ICSTSN53084.2022.9761294.

[5] Sohrabpour, V., Oghazi, P., Toorajipour, R., & Nazarpour, A. (2020). Export sales forecasting using artificial intelligence. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2020.120480.

[6] Penpece, D., Elma, E., & Adana Science and Technology University. (2014). Predicting Sales Revenue by Using Artificial Neural Network in Grocery Retailing Industry: A Case Study in Turkey. *International Journal of Trade, Economics and Finance, 5(5), 411-415*. http://ijtef.org/papers/411-E100.pdf "

[7] Huber, J., & Stuckenschmidt, H. (2020). Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, 36, 1420-1438. https://doi.org/10.1016/j.ijforecast.2020.02.005.

[8] Ahmedov, A. (2020). Walmart Sales Forecast. Kaggle. https://www.kaggle.com/datasets/aslanahmedov/walmart-sales-forecast"

[9] Luengo, J., García, S., & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, 32, 77-108. https://doi.org/10.1007/s10115-011-0424-2.

[10] Schober, P., Boer, C., & Schwarte, L. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126, 1763–1768. https://doi.org/10.1213/ANE.0000000000002864.

[11] Yusuf, A., & Alawneh, S. (2018). GPU Implementation of Sales Forecasting with Linear Regression. *CompSciRN: Computational (Topic)*. https://doi.org/10.2139/ssrn.3531033.

[12] Barrow, D., & Crone, S. (2016). Cross-validation aggregation for combining autoregressive neural network forecasts. *International Journal of Forecasting*, 32, 1120-1137. https://doi.org/10.1016/J.IJFORECAST.2015.12.011.

[13] Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2020). A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, 1-26. https://doi.org/10.1007/s40745-020-00253-5.

[14] Cherfi, S., Akoka, J., & Comyn-Wattiau, I. (2006). Use Case Modeling and Refinement: A Quality-Based Approach. , 84-97. https://doi.org/10.1007/11901181_8.

[15] Sharma, S. (2018). Research data analysis with Power BI. *International Journal of Advance Research, Ideas and Innovations in Technology*, 4, 316-319.

[16] dairu, X., & Shilong, Z. (2021). Machine Learning Model for Sales Forecasting by Using XGBoost. *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 480-483. https://doi.org/10.1109/ICCECE51280.2021.9342304.

[17] Kalla, R., Murikinjeri, S., & Abbaiah, R. (2020). An Improved Demand Forecasting with Limited Historical Sales Data. *2020 International Conference on Computer Communication and Informatics (ICCCI)*, 1-5. https://doi.org/10.1109/ICCCI48352.2020.9104090.

[18] Tuppi, L., Ollinaho, P., Ekblom, M., Shemyakin, V., & Järvinen, H. (2020). Necessary conditions for algorithmic tuning of weather prediction models using OpenIFS as an example. *Geoscientific Model Development*. https://doi.org/10.5194/egusphere-egu2020-4824.