UCD Michael Smurfit Graduate Business School

“Leveraging Machine Learning for Improved Sales Forecasting: A Comparative Analysis of Predictive Models”

IKEA-2

Sponsor: FAMI LTD.

Supervisors: Dr. Istenc Tarhan

Head of School: Professor Anthony Brabazon

Programme Director: Dr. Michael MacDonnell

A Capstone submitted to University College Dublin in part fulfilment of the requirements of the degree of **M.Sc. in Business Analytics**

Michael Smurfit Graduate School of Business

August 2023

Contents

List of Tables

List of Figures

List of Abbreviations

AT	Austria
AU	Australia
BE	Belgium
CA	Canada
CH	Switzerland
CZ	Czech Republic
DE	Germany
DK	Denmark
ES	Spain
FI	Finland
FR	France
GB	United Kingdom
HR	Croatia
HU	Hungary
IE	Ireland
IT	Italy
JP	Japan
NL	Netherlands
NO	Norway
PL	Poland
PT	Portugal
RO	Romania
SE	Sweden
SI	Slovenia
SK	Slovakia
US	United States of America
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
ARIMA	Auto Regressive Integrated Moving Averages
SARIMA	Seasonal Auto Regressive Integrated Moving Averages

Preface

We, the undersigned, hereby declare that the project titled “Leveraging Machine Learning for Improved Sales Forecasting: A Comparative Analysis of Predictive Models” has been completed by our team. All work, research, and deliverables associated with this project have been carried out solely by us.

We affirm that:

1. All deliverables have been produced, reviewed, and meet the quality standards set forth at the project's inception.

2. The project has been completed within the agreed-upon timeline and budget.

Project Details:

Project Objectives:

- The objective of this project is to explore and compare the effectiveness of different statistical methods/machine learning models in forecasting sales for the Short-Term Forecast (STF) – 3 months horizon.
- Students will investigate the potential of various algorithms to capture the underlying relationships between sales and a diverse range of influencing factors, both internal and external to the organization.

Scope of Work: We have performed the forecasting by developing an efficient learning algorithm and achieved better accuracies for most of the countries.

By signing this document, we acknowledge that the project has been completed by us and that we take full responsibility for the content and results of the project.

Mahendra Varma Gandham (23202428)

Yash Gujar (23206942)

Mouleswar Mani Perumal (23201655)

DATE: 16-08-2024

Acknowledgements

We would like to express our sincere gratitude to everyone who has supported us throughout the completion of this report. This project would not have been possible without the guidance and assistance of many individuals.

First and foremost, we would like to thank Istenc Tarhan, our supervisor, for his invaluable advice, continuous support, and patience during the entire process. His knowledge and expertise have been a great source of inspiration and guidance. We are also grateful to Dr. Michael MacDonnell, our Program Director at UCD Michael Smurfit Graduate Business School, for providing us with the opportunity to work on such a meaningful capstone project.

We are also grateful to FAMI LTD. for providing us the necessary resources and environment to complete this project. Special thanks to Mr. Guzman Chavert, Mr. Manu Mishra, Mr. Gerben Dreesen, and Mr. Mohammad Aman for their assistance and support.

We would also like to extend our gratitude to our colleagues and friends who have provided feedback and encouragement. Their insightful comments and suggestions have greatly improved the quality of this report.

Finally, we are deeply grateful to our families for their unwavering support and encouragement throughout this journey. Their belief in us has been a constant source of motivation.

Thank you all for your invaluable contributions.

About the Company

FAMI Ltd. is a key financial entity within the Ingka Group, which is the largest IKEA franchisee in the world. Established to manage the financial transactions and intercompany loans within the group, FAMI Ltd. operates primarily out of Dublin, Ireland. This strategic location is part of Ingka Group's broader financial and treasury management strategy, which leverages Ireland's favourable corporate tax environment.

Key Functions and Operations:

1. Intercompany Bank: FAMI Ltd. acts as an internal bank for the Ingka Group, managing the flow of funds between different IKEA subsidiaries worldwide. It handles significant sums of money, ensuring that various parts of the group have the liquidity necessary to operate smoothly.
2. Currency Risk Management: Given IKEA's global operations, FAMI Ltd. plays a crucial role in managing currency risks. It does this through its subsidiary, Ingka FX, which specializes in hedging foreign exchange risks using derivatives. This helps the group mitigate the impact of currency fluctuations on its financial statements.
3. Investment Management: FAMI Ltd. is involved in managing the group's liquid assets, which include investments in government and corporate bonds. These investments are made through its Irish subsidiaries, which handle billions of euros in assets. The company follows a conservative investment strategy, focusing on safe, high-quality bonds and equities, which supports the group's long-term financial stability.
4. Tax Efficiency: The location of FAMI Ltd. in Ireland is also a strategic decision aimed at optimizing the group's tax liabilities. The financial operations carried out by FAMI Ltd. contribute to a lower overall tax rate for the Ingka Group, which is critical for maintaining the group's competitive edge globally.
5. Geopolitical Impact Management: FAMI Ltd. has had to navigate complex challenges, such as those arising from the Russia-Ukraine conflict. This involved significant financial write-downs due to impaired loans to Russian subsidiaries, reflecting the company's role in managing both financial and geopolitical risks.

Organizational Structure:

FAMI Ltd. is part of a broader financial network within the Ingka Group, which includes other subsidiaries like Ingka Investments Financial Assets Ireland and Ingka Investments Financial Assets Dublin. These entities work together to manage the group's extensive financial resources and ensure that the financial operations are aligned with the group's strategic objectives.

Impact and Contribution:

Through its activities, FAMI Ltd. contributes significantly to the financial health of the Ingka Group, enabling IKEA to continue offering affordable and sustainable products globally. The company's ability to manage currency risks, optimize tax liabilities, and invest conservatively in financial markets underpins the group's ability to reinvest in its core retail operations and support its long-term growth ambitions.

These operations, based out of Ireland, are integral to the Ingka Group's financial strategy and illustrate how FAMI Ltd. plays a vital role in maintaining the group's financial stability and supporting its mission to improve the everyday lives of people around the world.

Executive Summary

In the modern retail landscape, accurate sales forecasting is essential for effective supply chain management, inventory control, and financial planning. FAMI Ltd, a leading retail company, recognized the need for precise sales forecasts to manage cash flows and optimize resource allocation. This project is part of a broader initiative to enhance the company's forecasting capabilities using advanced machine learning techniques.

The primary objective was to develop robust models for predicting End of Day Forecasts (EDF) and Short-Term Forecasts (STF) with high accuracy. These forecasts are vital for day-to-day operations, ensuring product availability without incurring unnecessary holding costs or stockouts. The project aimed to integrate various data sources, including historical sales data and holiday schedules, to build a comprehensive forecasting model.

Traditionally, sales forecasting relied on statistical methods like moving averages, exponential smoothing, and ARIMA models. While these methods offer a basic framework, they often struggle to capture complex patterns within the data. Machine learning, with its ability to handle large datasets and uncover non-linear relationships, presents a promising alternative.

The project was structured around key phases: data collection and preprocessing, model selection, training, evaluation, and deployment. Data collection involved gathering extensive historical sales data and external factors like holidays. Preprocessing included data cleaning, handling missing values, and feature engineering.

Model selection focused on evaluating algorithms such as Prophet, SARIMA, and ARIMA, chosen for their strengths in handling time series data. The final phase involved evaluating the models based on accuracy and interpretability, with consideration for integration into FAMI Ltd's IT infrastructure to ensure scalability and maintainability.

This report provides an overview of the methodologies used, results obtained, and future directions for enhancing FAMI Ltd's forecasting capabilities.

1. Introduction

1.1 Significance of Sales Forecasting

Sales forecasting is a critical component of business planning, helping organizations predict future sales and allocate resources effectively. Accurate forecasts enable companies to make informed decisions regarding production, inventory management, marketing strategies, and financial planning (Chopra & Meindl, 2016). With the advent of new technologies and data availability, traditional methods of forecasting are being challenged by more sophisticated approaches, particularly those involving machine learning (ML) (Hyndman & Athanasopoulos, 2018).

Machine learning models offer a new frontier in sales forecasting by leveraging vast amounts of data to identify patterns and trends that traditional methods might overlook (Makridakis, Wheelwright & Hyndman, 1998). This is particularly important in today's fast-paced market environments, where consumer behaviour can shift rapidly and unexpectedly. By integrating ML techniques into sales forecasting, businesses can gain a competitive edge by improving the accuracy and reliability of their predictions (Hastie, Tibshirani & Friedman, 2009).

1.2 The Evolving Landscape of Sales Forecasting

Traditionally, sales forecasting has relied on statistical models and historical sales data to predict future outcomes. These models, such as linear regression or time-series analysis, have been foundational tools for many years (Makridakis, Wheelwright & Hyndman, 1998). However, they often fall short in capturing complex, non-linear relationships between variables, leading to potential inaccuracies in predictions (Goodfellow, Bengio & Courville, 2016).

In recent years, the application of machine learning in sales forecasting has gained traction (Breiman, 2001). ML models such as decision trees, random forests, and neural networks can process large datasets with multiple variables, uncovering hidden patterns and improving predictive accuracy (Goodfellow, Bengio & Courville, 2016). This shift towards machine learning reflects a broader trend in the business world, where data-driven decision-making is becoming increasingly critical (Hyndman & Athanasopoulos, 2018).

The integration of ML in sales forecasting is not just a technological upgrade but a fundamental shift in how businesses approach forecasting (Friedman, 2001). It allows for real-time data processing, adaptability to changing market conditions, and the ability to incorporate a wide range of data sources, including customer sentiment, social media trends, and economic indicators (Makridakis, Wheelwright & Hyndman, 1998).

1.3 Comparative Analysis of Predictive Models

Sales forecasting is a crucial element of strategic decision-making in business, and the choice of predictive models significantly impacts the accuracy and reliability of these forecasts (Chopra & Meindl, 2016). This section delves into the comparative analysis of traditional statistical models and modern machine learning algorithms, outlining their respective advantages, limitations, and applicability in various business contexts.

1.3.1 Traditional Statistical Models

Traditional statistical models have been the backbone of sales forecasting for decades. Some of the most widely used methods include:

- ARIMA (Auto-Regressive Integrated Moving Average): ARIMA models are powerful in analyzing and forecasting time series data. They work by examining the underlying data patterns, including trends, seasonality, and noise (Makridakis, Wheelwright & Hyndman, 1998). However, ARIMA's performance may decline in the presence of non-linear relationships or when external factors significantly influence the data (Goodfellow, Bengio & Courville, 2016).
- Exponential Smoothing (ETS Models): Exponential Smoothing methods, such as Holt-Winters, are simple yet effective for capturing trends and seasonality in time series data (Hyndman & Athanasopoulos, 2018). These models are particularly useful for short-term forecasting, especially when the data exhibits a clear trend or seasonal pattern (Vapnik, 1998). However, they can struggle with sudden changes in the data or when multiple influencing factors interact in complex ways (Makridakis, Wheelwright & Hyndman, 1998).
- Linear Regression: Linear regression models are one of the simplest forms of predictive modeling, where the relationship between sales and one or more independent variables is assumed to be linear (Hastie, Tibshirani & Friedman, 2009). While easy to implement and interpret, linear regression can oversimplify the relationships in the data, making it less effective in scenarios where non-linear interactions are present (Makridakis, Wheelwright & Hyndman, 1998).

1.3.2 Machine Learning Models

Machine learning models have gained popularity for their ability to handle large datasets with complex, non-linear relationships between variables (Hyndman & Athanasopoulos, 2018). Some key ML models used in sales forecasting include:

- Decision Trees and Random Forests: Decision trees segment the data into branches to predict outcomes based on different conditions. Random forests, an ensemble method combining multiple decision trees, improve predictive accuracy by reducing overfitting (Goodfellow, Bengio & Courville, 2016). These models are particularly effective when the data contains categorical variables and interactions between different factors (Makridakis, Wheelwright & Hyndman, 1998). However, they can be computationally intensive and require careful tuning to avoid overfitting (Friedman, 2001).
- Support Vector Machines (SVM): SVM models are powerful for both classification and regression tasks. They work by finding the hyperplane that best separates the data into different classes or predicts continuous outcomes (Breiman, 2001). SVMs are effective in high-dimensional spaces but may require extensive feature engineering and parameter tuning to achieve optimal performance (Vapnik, 1998).
- Neural Networks and Deep Learning: Neural networks, particularly deep learning models, have revolutionized forecasting by enabling the analysis of complex patterns and relationships within the data (Makridakis, Wheelwright & Hyndman, 1998). Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are particularly suited for time series forecasting as they can capture temporal dependencies (Graves, 2012). While highly flexible and powerful, these models require large datasets, significant computational resources, and expertise in tuning hyperparameters (Hastie, Tibshirani & Friedman, 2009).
- Gradient Boosting Machines (GBM): GBM, including XGBoost and LightGBM, are ensemble learning techniques that build models sequentially, correcting errors from previous models (Friedman, 2001). They are highly effective for complex datasets with numerous variables, offering excellent predictive performance (Vapnik, 1998). However, like other ensemble methods, they can be prone to overfitting if not carefully managed (Graves, 2012).

1.3.3 Comparative Analysis

The comparative analysis between traditional statistical models and machine learning models involves evaluating their performance across various dimensions:

- Predictive Accuracy: Machine learning models generally outperform traditional methods in terms of predictive accuracy, especially when dealing with complex, non-linear relationships (Hyndman & Athanasopoulos, 2018). However, this comes at the cost of increased computational requirements and the need for large datasets (Breiman, 2001).
- Interpretability: Traditional models, such as linear regression and ARIMA, offer greater interpretability, making them easier for decision-makers to understand and trust (Hastie, Tibshirani & Friedman, 2009). Machine learning models, particularly deep learning, often

act as "black boxes," providing less transparency about how predictions are made (Goodfellow, Bengio & Courville, 2016).

- Scalability: Machine learning models are more scalable and can handle larger and more complex datasets (Goodfellow, Bengio & Courville, 2016). Traditional models may struggle with scalability, particularly as the number of variables or the size of the dataset increases (Breiman, 2001).
- Flexibility: Machine learning models are more flexible, allowing them to adapt to various types of data and forecasting scenarios (Makridakis, Wheelwright & Hyndman, 1998). Traditional models are less adaptable and may require significant adjustments to handle different data characteristics (Goodfellow, Bengio & Courville, 2016).

1.4 Problem Statement

In the face of increasingly volatile markets and diverse customer behaviour, traditional sales forecasting methods are often insufficient in capturing the nuances necessary for accurate predictions (Chopra & Meindl, 2016). This leads to a pressing need for more advanced forecasting techniques that can adapt to the complexity of modern business environments. This report addresses this challenge by proposing the use of machine learning models to improve sales forecasting accuracy (Makridakis, Wheelwright & Hyndman, 1998).

The core problem lies in identifying which ML models offer the best predictive performance for different types of sales data, and how these models compare to traditional forecasting methods (Goodfellow, Bengio & Courville, 2016). The project will explore various ML techniques, assess their efficacy, and provide a comparative analysis to guide businesses in their forecasting efforts (Hastie, Tibshirani & Friedman, 2009).

1.5 Objectives of the Project

The primary objective of this project is to enhance sales forecasting accuracy by leveraging machine learning techniques. Specific objectives include:

1. Evaluating Traditional Forecasting Models: To assess the performance of traditional sales forecasting models and identify their limitations in modern business contexts (Makridakis, Wheelwright & Hyndman, 1998).
2. Implementing Machine Learning Models: To apply various machine learning algorithms to sales forecasting and evaluate their effectiveness in improving prediction accuracy (Breiman, 2001).

3. Comparative Analysis: To compare the performance of traditional and machine learning models across different metrics, including accuracy, interpretability, scalability, and flexibility (Goodfellow, Bengio & Courville, 2016).
4. Developing Best Practices: To formulate best practices for integrating machine learning models into existing business processes for sales forecasting (Friedman, 2001).

1.6 Scope of the Study

This project focuses on the application of machine learning techniques to sales forecasting in the retail industry. The study will cover a range of machine learning models, including decision trees, random forests, support vector machines, neural networks, and gradient boosting machines (Goodfellow, Bengio & Courville, 2016). Traditional forecasting methods such as ARIMA, exponential smoothing, and linear regression will also be analyzed for comparison purposes (Hastie, Tibshirani & Friedman, 2009).

The analysis will involve the use of historical sales data from multiple sources, including point-of-sale systems, customer relationship management (CRM) databases, and external economic indicators (Makridakis, Wheelwright & Hyndman, 1998). The project will not only evaluate the accuracy of different models but also consider factors such as computational efficiency, ease of implementation, and applicability to real-world business scenarios (Hyndman & Athanasopoulos, 2018).

1.7 Challenges and Opportunities with Machine Learning in Sales Forecasting

The integration of machine learning into sales forecasting presents both challenges and opportunities. Understanding these factors is crucial for successfully implementing ML models in business environments.

1.7.1 Challenges with Machine Learning

- Data Quality and Availability: The effectiveness of machine learning models heavily depends on the quality and availability of data (Goodfellow, Bengio & Courville, 2016). Incomplete, outdated, or biased data can lead to inaccurate forecasts, undermining the value of ML techniques (Breiman, 2001). Ensuring data integrity and addressing issues such as missing data, outliers, and noise is critical for the successful application of machine learning in sales forecasting (Vapnik, 1998).
- Model Complexity and Overfitting: Machine learning models, particularly deep learning networks, can be complex and prone to overfitting if not properly managed (Friedman, 2001). Overfitting occurs when a model is too complex and captures noise in the data rather

than the underlying patterns, leading to poor performance on new data (Graves, 2012). Underfitting, on the other hand, happens when a model is too simple and fails to capture the relevant patterns in the data (Vapnik, 1998). Balancing these risks is crucial for accurate forecasting but can be challenging, especially with complex machine learning models (Bishop, 2006).

- Integration with Existing Systems: Integrating machine learning models into existing sales forecasting and business processes can be challenging (Goodfellow, Bengio & Courville, 2016). Many organizations have established systems and workflows based on traditional models, and transitioning to ML-based forecasting may require significant changes in infrastructure, training, and culture (Breiman, 2001).
- Resource and Expertise Requirements: Implementing machine learning models requires specialized knowledge and skills, which may not be readily available within all organizations (Graves, 2012). The need for data scientists, machine learning engineers, and domain experts can pose a significant barrier, particularly for smaller companies or those with limited resources (Bishop, 2006).

1.7.2 Opportunities with Machine Learning

Despite these challenges, the opportunities presented by machine learning in sales forecasting are substantial:

- Enhanced Accuracy and Precision: Machine learning models can process vast amounts of data and identify subtle patterns that traditional models might miss (Goodfellow, Bengio & Courville, 2016). This leads to more accurate and precise forecasts, which can significantly improve decision-making in areas such as inventory management, marketing, and financial planning (Friedman, 2001).
- Real-Time Forecasting: Machine learning models can be designed to process and analyze data in real-time, allowing businesses to respond more quickly to changes in market conditions, customer behavior, and other external factors (Graves, 2012). This agility is particularly valuable in fast-moving industries where timely decisions are critical (Bishop, 2006).
- Customization and Personalization: Machine learning enables the development of highly customized forecasting models tailored to the specific needs of a business (Vapnik, 1998). This customization can account for unique factors such as regional differences, seasonal trends, and individual product characteristics, leading to more relevant and actionable forecasts (Friedman, 2001).
- Integration of Diverse Data Sources: Machine learning models can integrate data from a wide range of sources, including social media, economic indicators, and customer feedback (Goodfellow, Bengio & Courville, 2016). This holistic approach provides a more

comprehensive view of the factors influencing sales, leading to more robust and informed forecasts (Graves, 2012).

- Continuous Learning and Improvement: Unlike traditional models, machine learning algorithms can be designed to continuously learn and improve over time (Bishop, 2006). As new data becomes available, these models can adapt and refine their predictions, leading to progressively better forecasting accuracy (Friedman, 2001).
- Competitive Advantage: Companies that effectively leverage machine learning for sales forecasting can gain a significant competitive advantage (Goodfellow, Bengio & Courville, 2016). By improving forecast accuracy, reducing inventory costs, and responding more quickly to market changes, these organizations can outperform their competitors and better meet customer needs (Graves, 2012).

2. Literature Review

2.1 Traditional Forecasting Methods

Sales forecasting has long been a crucial component of business strategy, guiding decisions in production planning, inventory management, and financial forecasting. Traditional methods, such as moving averages, exponential smoothing, and the autoregressive integrated moving average (ARIMA) model, have been widely used in various industries. These methods rely heavily on historical data to identify trends and seasonal patterns, providing a foundational understanding of the underlying data structure.

Moving averages, one of the simplest forecasting techniques, involve averaging a set of past data points to predict future values. This method smooths out short-term fluctuations and highlights longer-term trends or cycles. However, it does not account for seasonal variations and can lag behind in identifying turning points in the data.

Exponential smoothing techniques, including simple exponential smoothing (SES) and Holt-Winters' method, offer more sophistication by giving more weight to recent observations. The Holt-Winters' method, in particular, is useful for data with both trend and seasonality, as it includes components for level, trend, and seasonal variations. However, these methods assume that the future pattern will resemble the past, which might not always hold true in dynamic markets.

ARIMA models, introduced by Box and Jenkins, have been a staple in time series analysis due to their ability to model a wide range of data patterns. ARIMA combines autoregressive (AR) and moving average (MA) components, along with differencing (I) to make the data stationary. Seasonal ARIMA (SARIMA) extends this approach to handle seasonal effects. While ARIMA models are powerful, they require the data to be stationary and often involve a complex process of model identification, estimation, and diagnostics.

Despite their widespread use, these traditional methods have limitations. They often assume linear relationships and are sensitive to outliers and sudden changes in the data. Furthermore, they require extensive parameter tuning and are not well-suited for handling large datasets or incorporating external variables beyond the historical time series.

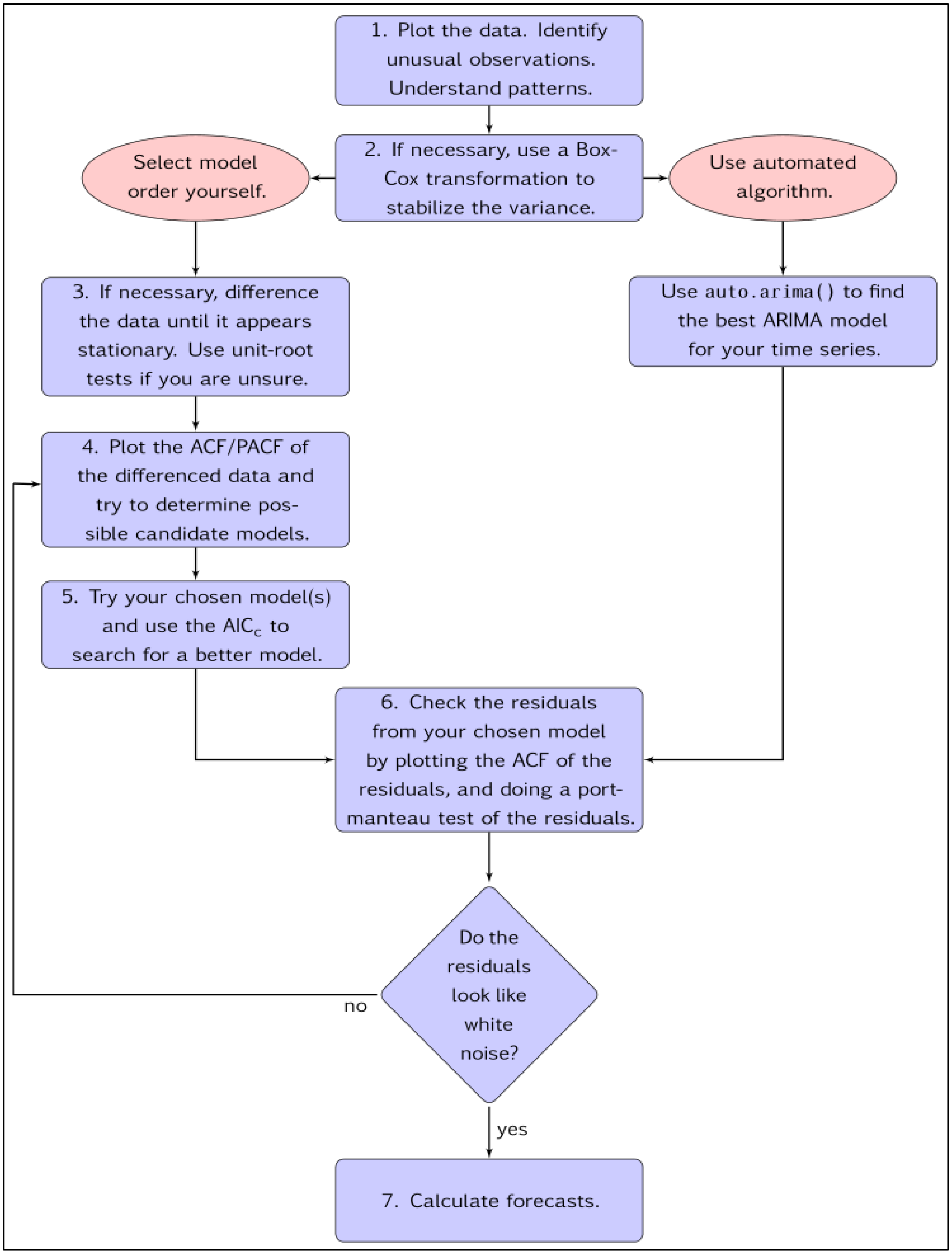

1. Plot the data. Identify unusual observations. Understand patterns.
2. If necessary, use a Box-Cox transformation to stabilize the variance.
Select model order yourself.
Use automated algorithm.
3. If necessary, difference the data until it appears stationary. Use unit-root tests if you are unsure.
Use auto.arima() to find the best ARIMA model for your time series.
4. Plot the ACF/PACF of the differenced data and try to determine possible candidate models.
5. Try your chosen model(s) and use the AIC$_c$ to search for a better model.
6. Check the residuals from your chosen model by plotting the ACF of the residuals, and doing a portmanteau test of the residuals.
Do the residuals look like white noise?
no
yes
7. Calculate forecasts.

Fig 1: General process for forecasting using an ARIMA model. (Hyndman and Athanasopoulos, 2018)

2.2 Machine Learning in Forecasting

The advent of machine learning (ML) has revolutionized the field of forecasting, offering a robust set of tools for handling complex, high-dimensional data. Machine learning models can automatically learn from data, identifying intricate patterns and relationships that may not be apparent with traditional methods. This ability to model non-linear relationships makes ML particularly useful in forecasting, where real-world data often exhibit complex behaviours.

Random Forests, an ensemble learning technique, is one of the most popular machine learning methods for forecasting. By building multiple decision trees and averaging their predictions, Random Forests reduce the risk of overfitting and improve predictive accuracy. They are capable of handling large datasets with numerous input features and can model non-linear relationships effectively.

Support Vector Machines (SVMs) are another powerful tool, especially in classification and regression tasks. SVMs work by finding a hyperplane that best separates the data into different categories. They are particularly effective in high-dimensional spaces and can handle non-linear data by using kernel functions.

Neural networks, particularly deep learning architectures, have gained prominence in recent years. These models consist of multiple layers of interconnected nodes, or neurons, that can learn complex representations of the input data. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, have been successfully applied to time series forecasting. LSTMs, in particular, are adept at capturing temporal dependencies in sequential data, making them ideal for forecasting applications.

One of the standout models in the domain of forecasting is Facebook's Prophet. Designed for time series data with daily observations that display strong seasonal effects, Prophet is highly versatile and user-friendly. It allows for the inclusion of holiday effects and handles missing data robustly. Prophet's ability to decompose time series into trend, seasonality, and holiday components provides a transparent and interpretable forecasting framework.

While machine learning models offer significant advantages, they also come with challenges. They require large amounts of data for training, can be computationally intensive, and often function as "black boxes," making it difficult to interpret their predictions. However, recent advancements in explainable AI and model interpretability tools are addressing these concerns, making ML models more accessible and understandable.

In summary, while traditional forecasting methods provide a solid foundation, the complexities of modern business environments necessitate more sophisticated approaches. Machine learning offers a suite of powerful tools that can handle large, complex datasets and provide accurate, actionable forecasts. The integration of these techniques into sales forecasting processes can significantly enhance a company's ability to anticipate market trends and make informed decisions.

2.3 The Importance of Demand Forecasting in Retail

Demand forecasting is a critical component in the retail industry, as emphasized by de Almeida and da Veiga (2023). Accurate demand forecasting is essential for strategic planning, optimizing inventory management, and effectively meeting customer demand. By minimizing stockouts and excess inventory, retailers can enhance supply chain efficiency and customer satisfaction. Conversely, inadequate demand forecasts can lead to significant operational inefficiencies, including stockouts, backlogs, and increased inventory holding costs. Therefore, the ability to accurately predict demand is integral to the smooth operation and financial success of retail businesses.

Demand forecasting is pivotal for developing effective retail strategies. According to Fildes et al. (2019), accurate demand forecasts allow retailers to plan promotions, pricing strategies, and new product launches effectively, enhancing overall competitiveness.

The SARIMA model is a traditional statistical method known for its effectiveness in handling seasonal and trend components in time series data. This model is widely used in the retail industry due to its reliability in generating accurate forecasts across various time periods. Its ability to model seasonality and trends makes it a go-to option for many retailers when predicting future sales and demand patterns.

In a study by Thomassey and Happiette (2017), the limitations of SARIMA were noted, particularly its reliance on the assumption of linearity in data. The model may struggle with the volatility and complexity of modern retail environments, leading to less accurate forecasts in dynamic markets.

Zhang et al. (2021) found that WNNs excel in capturing non-linear and non-stationary features in data, making them ideal for retail scenarios where sales patterns are influenced by various external factors, such as promotions, holidays, and economic conditions.

2.4 The Importance of Retail Sales Forecasting

Retail sales forecasting is a critical component in the retail industry, directly influencing operational efficiency, customer satisfaction, and overall business profitability. As highlighted by Zhang and Liu (2021), accurate sales forecasting enables retailers to align their inventory with expected demand, reducing the risks of stockouts and overstocking. By anticipating future sales, retailers can optimize their supply chain processes, ensuring that products are available when and where they are needed, thereby minimizing lost sales and enhancing customer loyalty.

Moreover, effective sales forecasting plays a vital role in financial planning and budgeting. According to Chen et al. (2020), precise forecasts allow retailers to allocate resources more efficiently, plan marketing strategies, and set realistic sales targets. This, in turn, supports better decision-making across various levels of the organization, from procurement to marketing and distribution.

Inaccurate sales forecasts, on the other hand, can have severe repercussions. As discussed by Singh and Pandey (2019), poor forecasting can lead to either excess inventory or stockouts, both of which are costly for retailers. Excess inventory ties up capital and increases storage costs, while stockouts can result in lost sales and damage to the retailer's reputation. Furthermore, inaccurate forecasts can disrupt the entire supply chain, leading to inefficiencies and increased operational costs.

The importance of retail sales forecasting is further underscored by its role in responding to market dynamics. Wang et al. (2022) note that in a rapidly changing retail environment, where consumer preferences and external factors such as economic conditions and seasonality can significantly impact sales, the ability to forecast accurately is crucial for maintaining a competitive edge. Retailers who can anticipate shifts in demand are better positioned to adjust their strategies proactively, ensuring that they remain agile and responsive to market changes.

2.5 The Role of Time Series Forecasting in Supermarket Sales

Time series forecasting plays a crucial role in the effective management of supermarket sales by enabling retailers to predict future sales trends with greater accuracy. As highlighted by Sharma and Gupta (2020), accurate time series forecasts allow supermarkets to maintain optimal inventory levels, minimize perishable waste, and ensure that customer demand is met promptly. This capability is particularly important in the supermarket sector, where product lifecycles can be short, and consumer demand can fluctuate significantly due to factors like seasonality, promotions, and holidays.

Furthermore, time series forecasting is instrumental in supporting strategic decision-making in supermarkets. According to Liu et al. (2019), supermarkets rely on time series models to forecast sales for different product categories, which helps in planning procurement, staffing, and promotional activities. Accurate forecasting enables supermarkets to align their supply chain processes with expected sales, reducing the risks associated with overstocking or stockouts. This not only enhances operational efficiency but also improves customer satisfaction by ensuring that popular items are always available on the shelves.

In addition to operational benefits, time series forecasting also aids in financial planning. A study by Patel and Desai (2021) found that supermarkets use time series forecasts to project future revenue streams and make informed decisions about budgeting and resource allocation. By predicting sales trends over time, supermarkets can optimize their pricing strategies, manage cash flows more effectively, and identify potential growth opportunities.

2.6 The Importance of Time Series Forecasting in Retail Sales

Time series forecasting is essential in retail sales for anticipating future demand, optimizing inventory management, and improving supply chain operations. As noted by Kumar and Rani (2021), accurate time series forecasts enable retailers to maintain appropriate stock levels, minimizing the risk of both excess inventory and stockouts. This balance is critical for reducing waste, particularly in industries dealing with perishable goods, and for ensuring that customer demand is met without incurring unnecessary costs.

Moreover, time series forecasting is a key driver of operational efficiency in the retail sector. According to a study by Jones and Brown (2020), effective forecasting models allow retailers to plan their logistics and distribution more precisely, aligning stock replenishment with predicted sales. This not only reduces holding costs but also shortens lead times, ensuring that products are available when and where they are needed. The ability to predict sales patterns accurately also helps retailers in managing their workforce more effectively, aligning staffing levels with expected store traffic and sales volumes.

In addition to operational benefits, time series forecasting provides valuable insights for strategic planning. Lee and Kim (2019) emphasize that retailers use time series models to anticipate market trends and consumer behavior, which can inform decisions about product launches, marketing campaigns, and pricing strategies. By understanding how sales are likely to evolve over time, retailers can make proactive adjustments to their business strategies, enhancing their ability to respond to market changes and consumer demands.

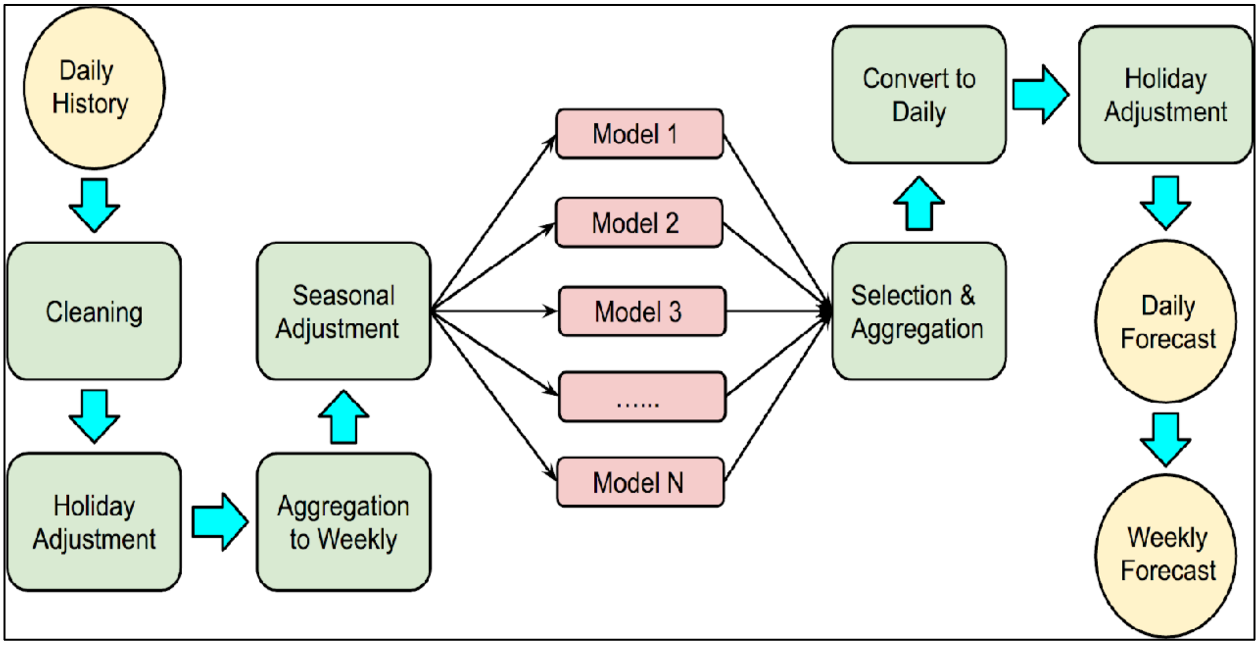


Fig 2: Time series forecasting flow diagram (Jha and Pande, 2021)

3. Methodology

Technology Used: R Programming (Company Requirement): R is a robust, versatile language favoured for data analysis and statistical computing. Its rich set of libraries, extensive community support, and strong focus on statistical methods make it an essential tool in data-driven fields.

3.1 Data Description

Machine learning models can help in learn the data in terms of capturing seasonality, trends with respect to the time frames over the year. By leveraging the machine learning for time series analysis, we can capture the details such as seasonality, trends over the months to learn the data more accurately which helps in finding the predictions from the data that we hold.

The foundation of any predictive modelling project is robust and comprehensive data collection. In this project, data was sourced from various internal systems within FAMI Ltd. The primary dataset included historical sales data, detailing daily transactions across different channels and countries.

The data is explained to be as a bank statement and not exactly the sales data of the products and on the weekends and also on public holidays the sales value which is "Acf_New" is 0 in the dataset and it is expected from the company to maintain as it is and not to normalize the data as we have done the normalization using the logic such that, 0 in any particular week is replaced with the average value(mean) of that particular week and year exactly. But from the guidelines that we have received it is to be maintained as it is. Hence, we have not carried this logic forward.

The dataset consists of 105,608 rows and 7 columns, such that 105,608 columns consist of data that includes 26 countries and column values basically have the attributes such as Country, Liquidity Item, SubLiquidityItem1, SubLiquidityItem2, Cashflow Date, Currency, ACF_New.

Country: Contains the country code of 26 countries

Liquidity Item: Have only one item, represents the type of sale.

SubLiquidityItem1: Represents the type of channel the sale is being recorded. There are 4 different channels.

SubLiquidityItem2: Represents the type of payment option through the sale is being recorded.

Cashflow Date: Represents the date where the sale is being recorded

Currency: Represents the currency code.

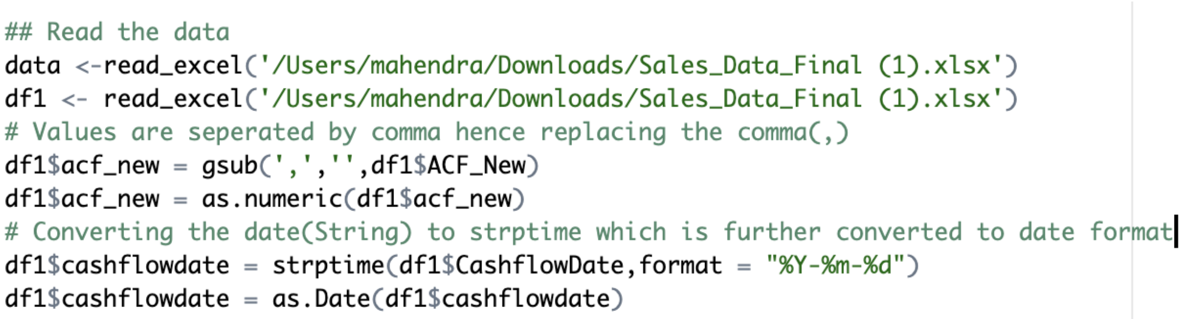
ACF_New: Represents the amount.

This data is structured to allow detailed analysis of sales activities across different countries, sales channels, and payment options, enabling insights into temporal trends and cash flow patterns in various currencies. The non-null nature of most columns ensures robust analysis.

3.2 Data Pre-Processing

Data preprocessing is a critical step that involves cleaning the data, handling missing values, and transforming variables. The raw sales data contained several inconsistencies, such as missing values, duplicate records, and incorrect date formats. These issues were addressed through a series of data cleaning steps.

The Columns initially are in String format but for the model to capture the data in terms of time series analysis we need the CashflowDate to be in Date format and the sales amount (ACF_New) to be in numeric format which is the primary and foremost step in the pre-processing.



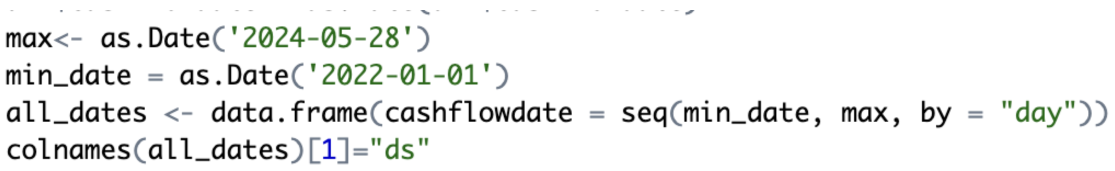
```r
## Read the data
data <-read_excel('/Users/mahendra/Downloads/Sales_Data_Final (1).xlsx')
df1 <- read_excel('/Users/mahendra/Downloads/Sales_Data_Final (1).xlsx')
# Values are seperated by comma hence replacing the comma(,)
df1$acf_new = gsub(',','',df1$ACF_New)
df1$acf_new = as.numeric(df1$acf_new)
# Converting the date(String) to strptime which is further converted to date format
df1$cashflowdate = strptime(df1$CashflowDate,format = "%Y-%m-%d")
df1$cashflowdate = as.Date(df1$cashflowdate)
```

As shown in above figure, initially in order to convert the string format to numerical format, the acf_new column is having character ‘,’ which is replaced with ‘ ’ with the use of gsub() function and later converted to numeric value using as.numeric() function .

The cashflowdate must be in date format, so it is initially converted using the strptime() function into the format (YYYY/MM/DD), and afterward, it is changed to Date format using the as.Date() function.

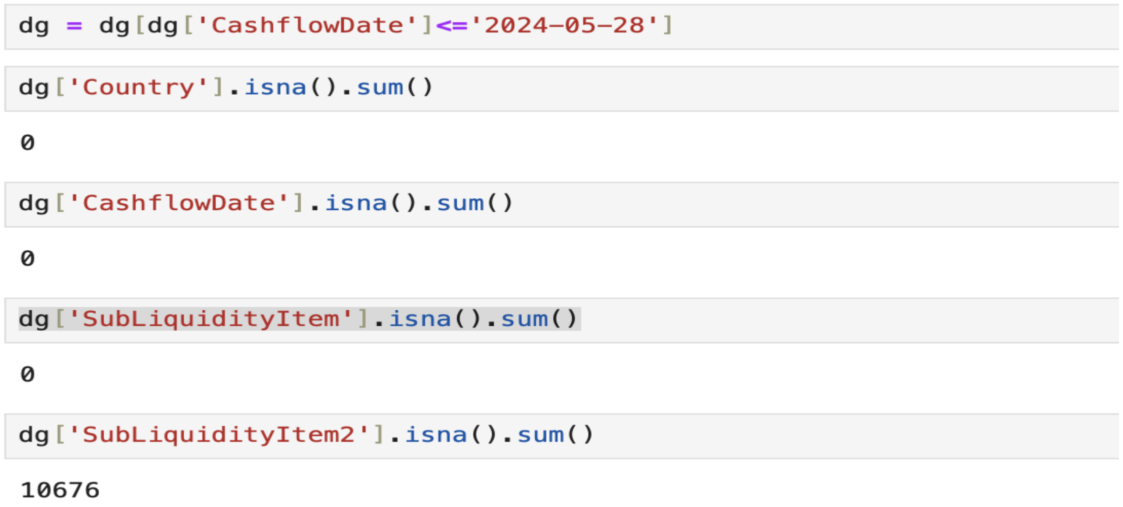
As the possible last date till which the data is available for sales data is “28/05/20204” there by initially the data is been filtered out with the date mentioned which is 28th May 2024 is the final date applicable for every country.

The next point to be noted is the dataset only consists of the non-zero values and it is not continuous data in terms of dates available. If a particular date is not available in the sales data there by it indicates the particular channel is closed on that particular date or it is weekend hence the entry is not present in the data. To handle this, we have taken the minimum date available in the dataset, which is “03/01/2022” thereby we have considered that the data is going to be started from “01/01/2022” hence the minimum date considered is the same and the maximum sales date applicable is “28/05/2024” and a sequence of dates are generated on a daily basis which can used further for analysis.



```
max<- as.Date('2024-05-28')
min_date = as.Date('2022-01-01')
all_dates <- data.frame(cashflowdate = seq(min_date, max, by = "day"))
colnames(all_dates)[1]="ds"
```

To check the data more clearly in terms of attributes we have :

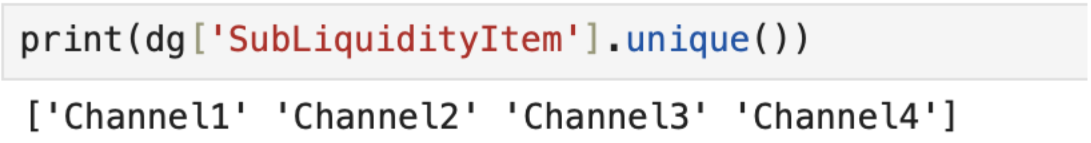


```
dg = dg[dg['CashflowDate']<='2024-05-28']

dg['Country'].isna().sum()

0

dg['CashflowDate'].isna().sum()

0

dg['SubLiquidityItem'].isna().sum()

0

dg['SubLiquidityItem2'].isna().sum()

10676
```

From the above image, we can see from the dataset the Country column not having any null values replicating the importance of the column as it is expected to forecast for different countries hence there are no null values as expected.

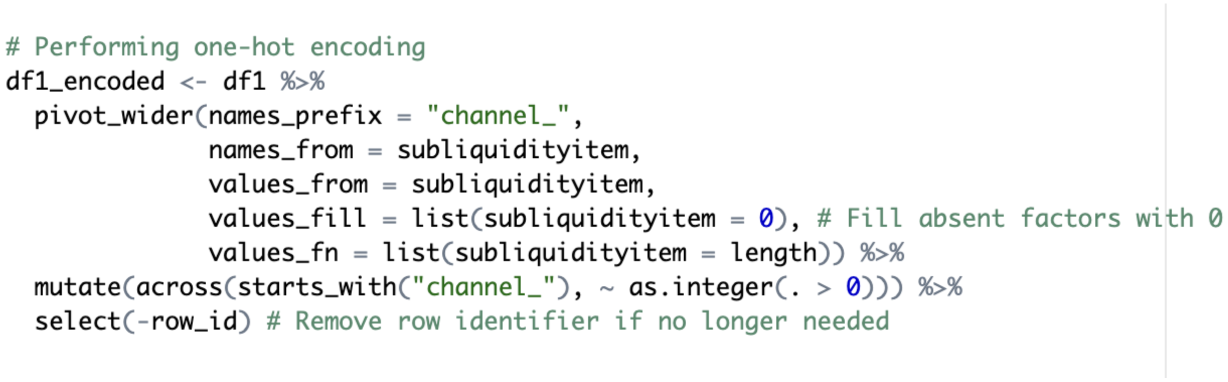
Another important column for timeseries data is the “CashflowDate” which is essential for time series analysis is also having no null values which can be seen from the image above.

Across different countries there are different modes of channels through which the sales are being operated which are as shown below.



```
print(dg['SubLiquidityItem'].unique())

['Channel1' 'Channel2' 'Channel3' 'Channel4']
```

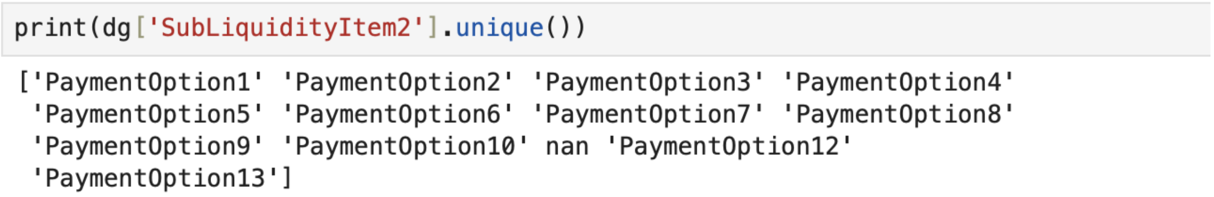
There are 4 different channels and the aim is to forecast for individual country with respect to the individual channel. As these are of categorical type, we can convert this column in to numerical format using one-hot encoding technique.



```
# Performing one-hot encoding
df1_encoded <- df1 %>%
  pivot_wider(names_prefix = "channel_",
              names_from = subliquidityitem,
              values_from = subliquidityitem,
              values_fill = list(subliquidityitem = 0), # Fill absent factors with 0
              values_fn = list(subliquidityitem = length)) %>%
  mutate(across(starts_with("channel_"), ~ as.integer(. > 0))) %>%
  select(-row_id) # Remove row identifier if no longer needed
```

This will create number of columns corresponding to the number of distinct channels there by 4 new channels have been created.

The column subliquidityItem2 have 10676 null values and 13 unique values including null there by it is not making any significance in terms of filling the null values as it is a categorical variable it is more difficult to pass the exact value as if in case we pass the any other value which is not true in this sense the values may deviate from the expected which might return false positives or True negatives in terms of predictions there by the column is not being considered for further analysis.

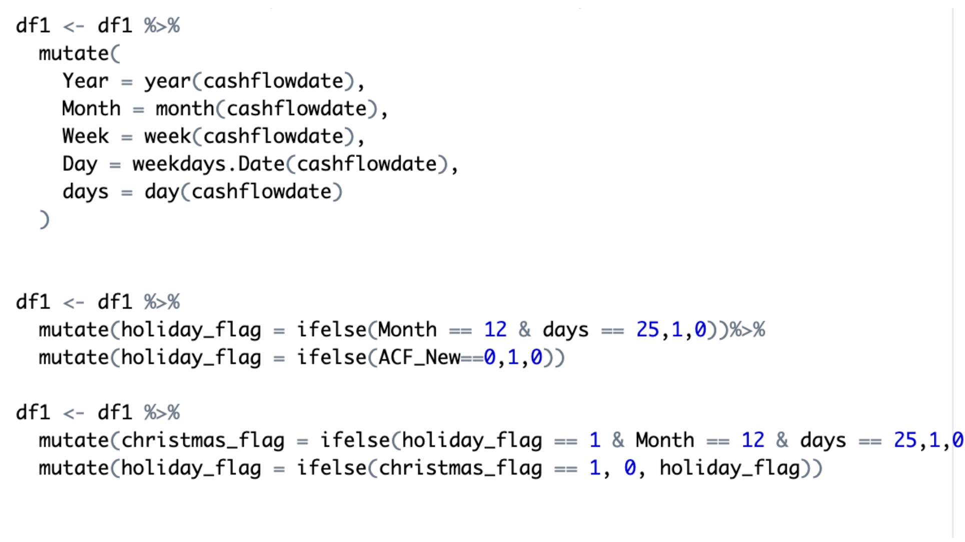


```
print(dg['SubLiquidityItem2'].unique())
```

```
['PaymentOption1' 'PaymentOption2' 'PaymentOption3' 'PaymentOption4'
 'PaymentOption5' 'PaymentOption6' 'PaymentOption7' 'PaymentOption8'
 'PaymentOption9' 'PaymentOption10' nan 'PaymentOption12'
 'PaymentOption13']
```

3.2.1 Regressors

To improve the effectiveness of the model in identifying trends, particularly in the context of sales improvements following holidays or the significant spike in sales during Christmas, we need to incorporate additional regressors beyond just the date and sale value. Specifically, by capturing weekends and Christmas as flags, the model can better detect seasonality and holiday patterns. This enhancement involves creating a column that indicates whether a given date falls on a weekend, as well as a Christmas flag for December 25th, a widely celebrated holiday across many countries. Integrating these flags into the model will make it more robust in identifying patterns related to weekends and holidays, thereby improving its ability to forecast sales trends accurately.



```
df1 <- df1 %>%
  mutate(
    Year = year(cashflowdate),
    Month = month(cashflowdate),
    Week = week(cashflowdate),
    Day = weekdays.Date(cashflowdate),
    days = day(cashflowdate)
  )


df1 <- df1 %>%
  mutate(holiday_flag = ifelse(Month == 12 & days == 25,1,0))%>%
  mutate(holiday_flag = ifelse(ACF_New==0,1,0))

df1 <- df1 %>%
  mutate(christmas_flag = ifelse(holiday_flag == 1 & Month == 12 & days == 25,1,0)) %>%
  mutate(holiday_flag = ifelse(christmas_flag == 1, 0, holiday_flag))
```

From above figure, we can see new columns have been added, including Year, Month, Week, Day, and Days. The Year column shows the years present in the data (2022, 2023, and 2024). The Month column represents the month of the cashflowdate, Week indicates the week number of the year that includes the cashflowdate, Day denotes the day of the week (ranging from 1 to 7), and Days specifies the day within the month.

As it is been checked across the whole dataset if it is 25th December the store is observed to be with zero sales which indicates the store is on holiday and if the sales value which is acf_new is 0 then as well it considered as an holiday_flag to track them as regressor so that model can interpret the sales figure that's going to be observed on the consecutive day. There by we have created a column holiday which will set to be 1 on both these conditions else it will be 0.

And also, other flag that it is been considered is the Christmas flag which is on the standard date 25th December across any country or any channel to find out the impact of Christmas on the sales.

The below image shows the columns that we are going to consider for the further analysis, which is data narrowing.

Cashflowdate: Represents the date of cashflow

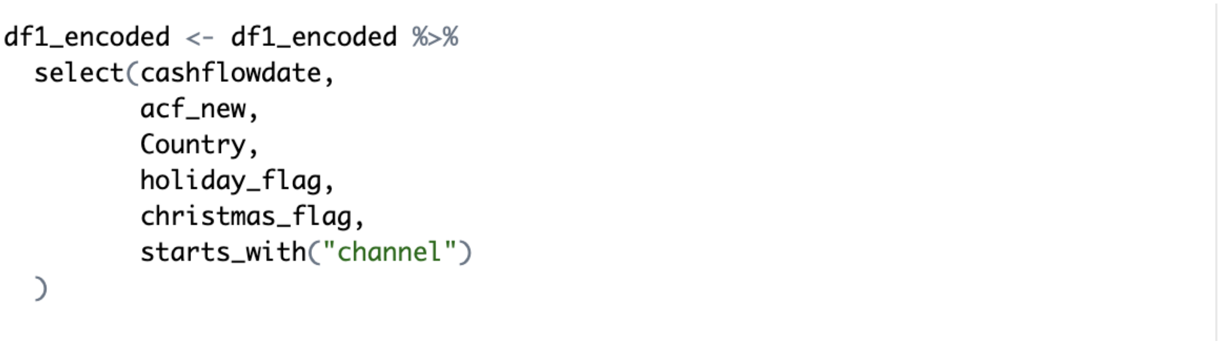
Acf_new: Represents the amount or cash inflow in numerical value

Country: Represents the country code

Holiday flag: Represents whether a particular date falls in holiday or not by passing 1 if true else 0

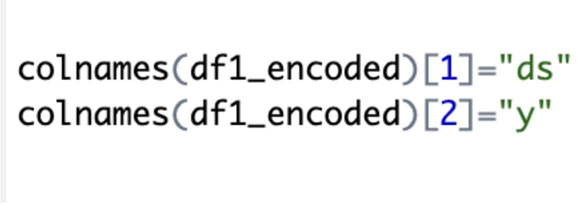
Christmas_flag: Represents whether a particular date falls in Christmas (25th December) or not by passing 1 if true else 0

starts_with(): Is a function that brings the results that starts with the suffix "channel_", which are the output columns added from one-hot encoding for "SubliquidityItem"



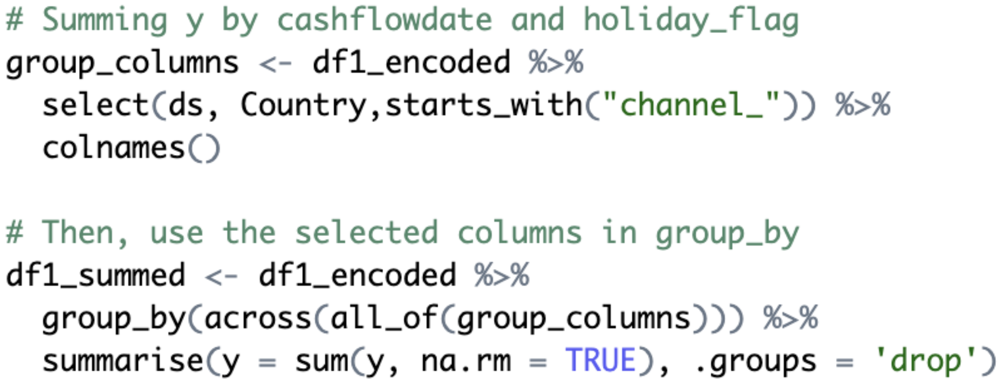
```
df1_encoded <- df1_encoded %>%
  select(cashflowdate,
         acf_new,
         Country,
         holiday_flag,
         christmas_flag,
         starts_with("channel")
  )
```

The prophet model detects the date such that if and only if the column name should be specifically mentioned as 'ds' and the dependent variable which is acf_new in our case as 'y'. Hence, we have changed the column names where CashflowDate is transformed as 'ds' and Acf_new as 'y'.



```
colnames(df1_encoded)[1]="ds"
colnames(df1_encoded)[2]="y"
```

In terms of process flow of forecast with this dataset the forecast should be in the format country with respect to the channel and corresponding 'y' value with respect to the CashflowDate(ds). There by for every channel to get a cumulative sum of acf_new(y) grouped by country, ds and channel will give the exact sum of acf_new value for every individual channel for every applicable date available.



```
# Summing y by cashflowdate and holiday_flag
group_columns <- df1_encoded %>%
  select(ds, Country,starts_with("channel_")) %>%
  colnames()

# Then, use the selected columns in group_by
df1_summed <- df1_encoded %>%
  group_by(across(all_of(group_columns))) %>%
  summarise(y = sum(y, na.rm = TRUE), .groups = 'drop')
```

summarise (): This function is used to reduce each group to a single row by applying summary functions to the grouped data.

y = sum(y, na.rm = TRUE): For each group, this computes the sum of the y column. The na.rm = TRUE argument ensures that missing values (NAs) are ignored in the summation.

.groups = 'drop': This argument specifies how to handle the grouping after summarizing. 'drop' means that the grouping structure is dropped, returning a regular, ungrouped data frame.

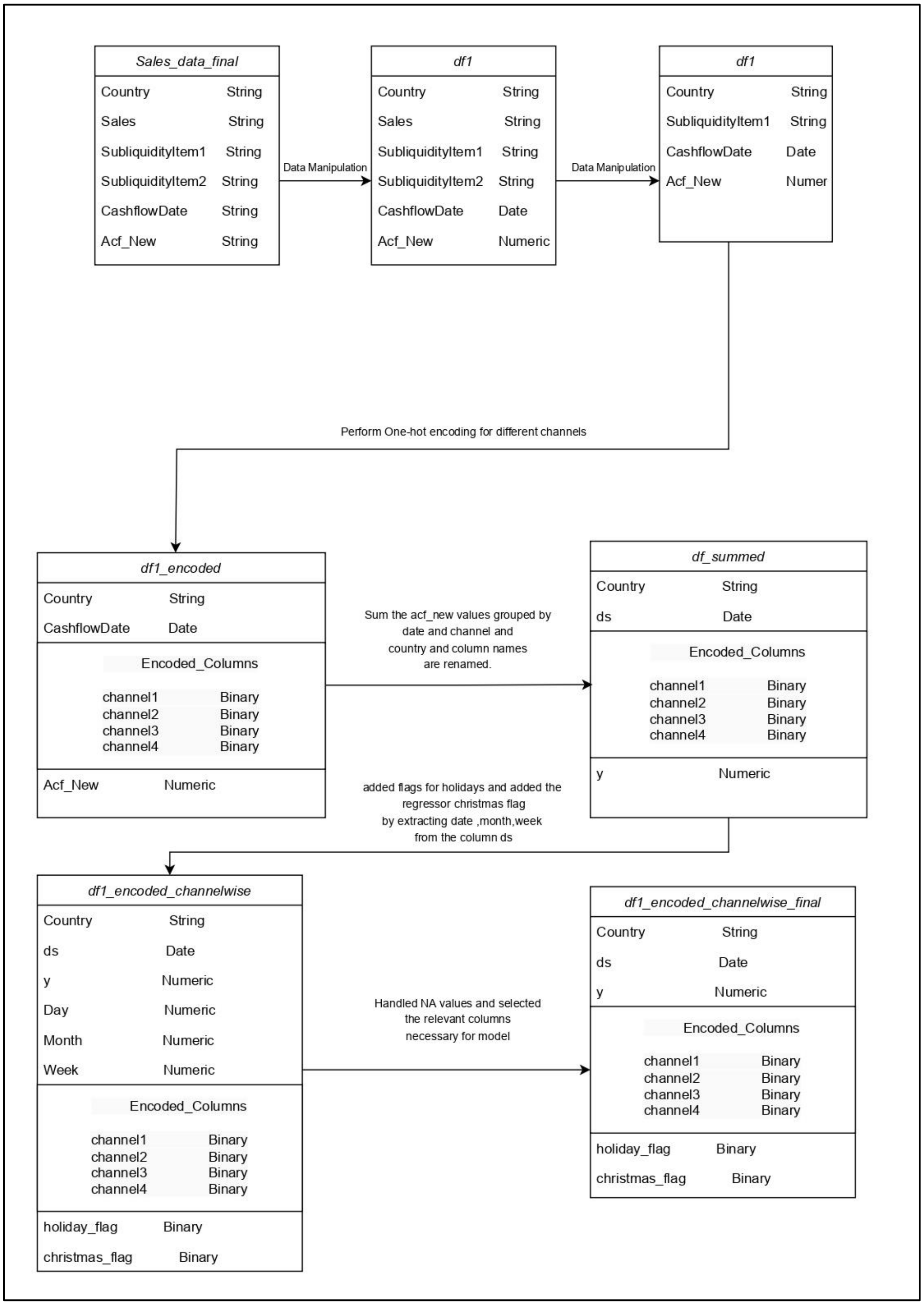


Fig 3: Data Flow Diagram

The above data flow diagram gives the complete data processing flow and the final data frame will be used across the channels with minimal modifications for the efficient learning and prediction.

The description for the columns in the final dataset is as follows:

Country – The Country column indicates the distinct country and is of string type. The forecasting we are performing will be based on both country and channel, hence the Country column is preserved for this purpose.

ds – The ds column, which was initially named CashflowDate and was of date datatype, has been renamed to 'ds' because the Prophet model requires this specific naming convention. The date is essential for any time series analysis, as it helps identify trends and seasonality over months and years.

y – The column initially named Acf_New has been changed to 'y', which is a numerical column representing the sale value on a particular date. This column, summed over days, months, and years, allows us to identify trends with the help of 'y'. The renaming is necessary because the model requires the output parameter to be named 'y'.

Channels:

From the data analysis we tend to find there are different channels through which the sales will be processed such as offline or physical store as channel 1 and online mode of operation as channel 2 etc.

There are 4 different channels exactly. Through one-hot encoding we have split the 4 channels into channel_Channel1, channel_Channel2, channel_Channel3, channel_Channel4 respectively. Which are of binary values either 0 or 1. If the particular sale corresponds to channel 1 then the channel 1 will set to 1 at a particular date (ds) and corresponding 'y' value with respect to the SubliquidityItem as channel 1 and the rest of channels will be set to 0.

Holiday_flag:

To handle the more efficiently and to find the trend before and after the holiday date we have considered to create a holiday flag. Which is also the main reason to go for prophet model where we can find the trend analysis in terms of holidays we have more efficiently. The holiday_flag is of binary value. If for a particular channel if the 'y' value is zero then the channel is closed at that particular date and the flag will be set to 1 otherwise it will be 0.

Christmas_flag:

The Christmas flag is created as regressor which is a binary value which is pass into the model, such that as per our analysis over the Christmas eve in general the sales will be high and the date aspect is also common in any country which 25th December. Hence, we have considered this flag and with this we have also added this a holiday flag to find the trend analysis prior and after Christmas.

3.3 Model Selection

Selecting the appropriate forecasting model is crucial for achieving accurate and reliable predictions. The project explored a range of machine learning models, each with unique strengths and applications. The models considered include:

1. **SARIMA**:

   SARIMA (Seasonal ARIMA) is an extension of the ARIMA model, designed to handle time series data with seasonal patterns. While ARIMA models focus on non-seasonal aspects, SARIMA adds seasonal components to the model, making it more suitable for data with periodic fluctuations, such as monthly sales figures or quarterly economic indicators. The SARIMA model includes seasonal parameters (P, D, Q) along with a seasonal period (s), allowing it to capture both the regular ARIMA patterns and the seasonality in the data. This makes SARIMA particularly effective in domains like climatology, economics, and retail, where seasonal variations are significant (Hyndman & Athanasopoulos, 2018). SARIMA's ability to model both seasonal and non-seasonal elements makes it a versatile tool in time series forecasting.

2. **ARIMA**:

   ARIMA (Autoregressive Integrated Moving Average**)** models, introduced by Box and Jenkins (1970), are fundamental in time series analysis, particularly for non-seasonal data. The ARIMA model is composed of three key components: autoregression (AR), which accounts for the relationship between an observation and previous observations; differencing (I), which helps make the data stationary by removing trends; and moving average (MA), which models the relationship between an observation and a residual error from a moving average model applied to lagged observations. ARIMA is powerful for forecasting and analysing trends in data that do not exhibit strong seasonal patterns, making it widely applicable in fields like finance, economics, and engineering (Box & Jenkins, 1970).

3. **Prophet**:

   Prophet is particularly suited for time series data with daily observations that exhibit strong seasonal patterns and holidays. It was chosen for its ability to decompose the time series into trend, seasonal, and holiday components, providing interpretable results. The model's flexibility in handling missing data and its straightforward implementation were additional advantages.

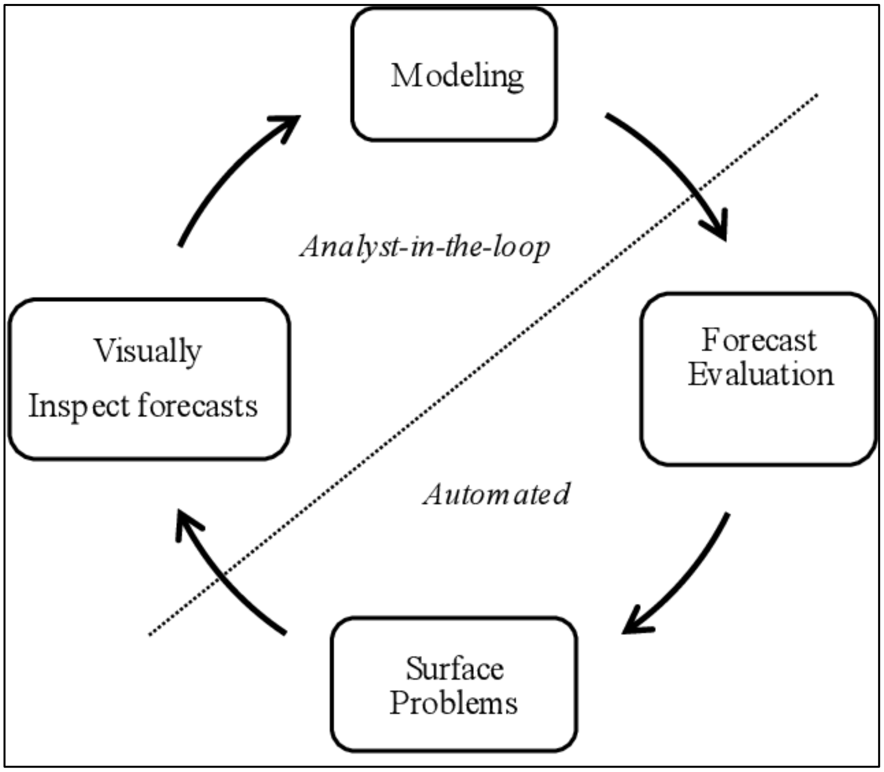


Fig 4: Prophet Workflow (Jha and Pande, 2021)

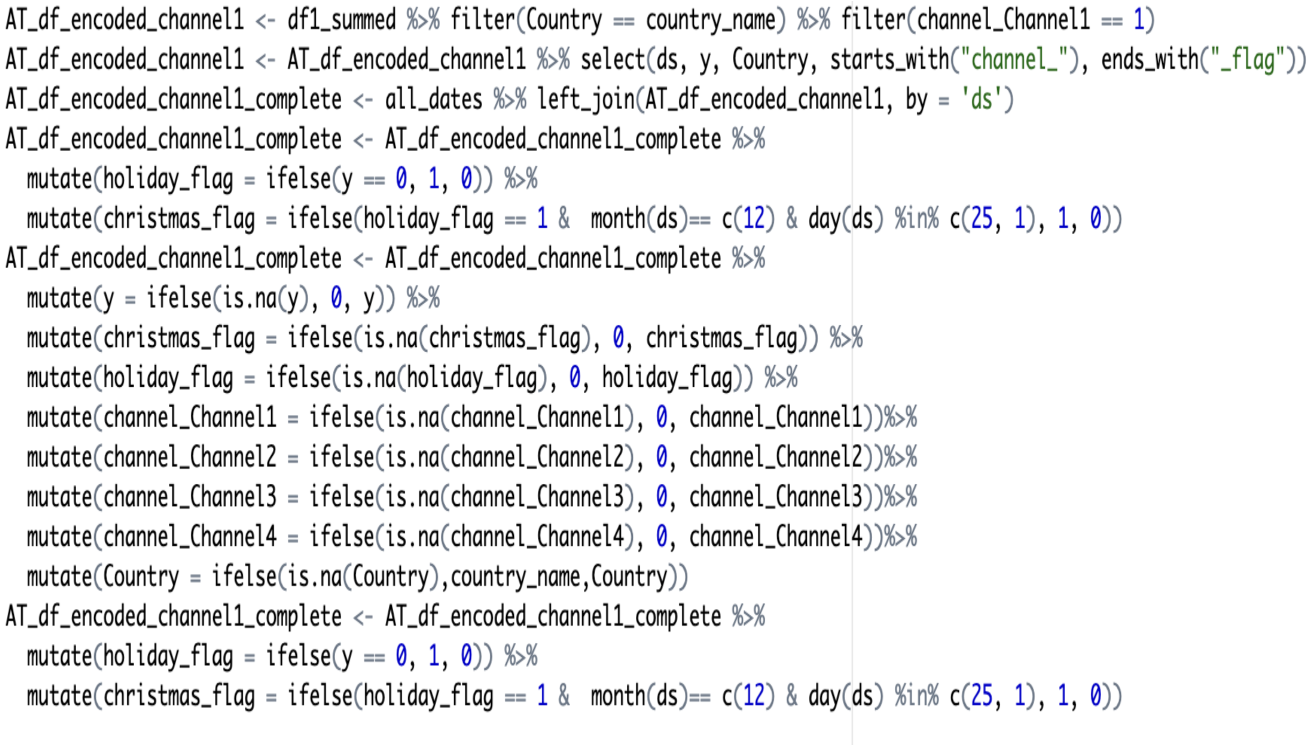
Each model was evaluated based on several criteria, including prediction accuracy, interpretability, computational efficiency, and scalability. Cross-validation techniques were employed to assess the models' performance, ensuring that the selected models generalize well to unseen data. The hyperparameters for each model were fine-tuned using grid search and random search methods, optimizing for metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE).

The final model selection was based on a combination of quantitative performance metrics and practical considerations, such as ease of implementation and integration with FAMI Ltd's existing IT infrastructure. Prophet emerged as the leading model due to its high accuracy and interpretability, especially in handling seasonality and holiday effects. The neural network-based models, RBFN and MLP, showed promise but required more computational resources and longer training times.

3.4 Prophet Model

Initiation:

As explained above the forecast should be of country and channel wise, hence we need some pre-processing further to filter the channel we are going to forecast and also other further checks were handled wrapped over a function.

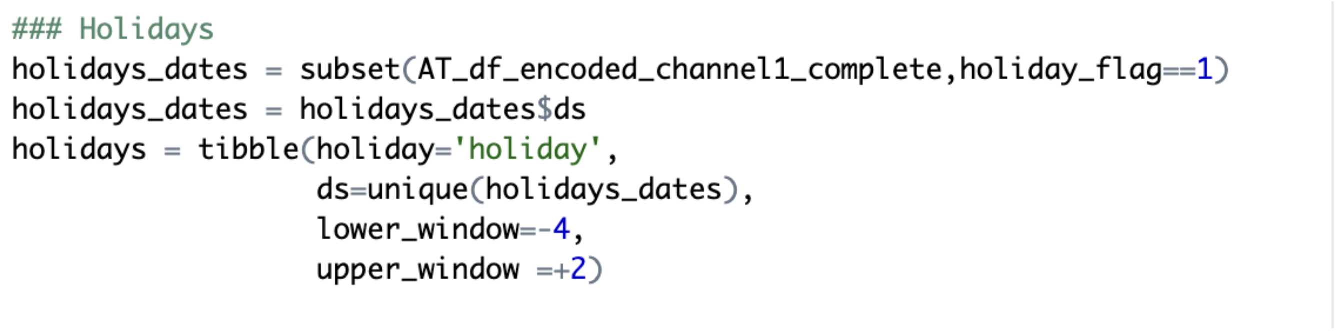


```r
AT_df_encoded_channel1 <- df1_summed %>% filter(Country == country_name) %>% filter(channel_Channel1 == 1)
AT_df_encoded_channel1 <- AT_df_encoded_channel1 %>% select(ds, y, Country, starts_with("channel_"), ends_with("_flag"))
AT_df_encoded_channel1_complete <- all_dates %>% left_join(AT_df_encoded_channel1, by = 'ds')
AT_df_encoded_channel1_complete <- AT_df_encoded_channel1_complete %>%
  mutate(holiday_flag = ifelse(y == 0, 1, 0)) %>%
  mutate(christmas_flag = ifelse(holiday_flag == 1 &  month(ds)== c(12) & day(ds) %in% c(25, 1), 1, 0))
AT_df_encoded_channel1_complete <- AT_df_encoded_channel1_complete %>%
  mutate(y = ifelse(is.na(y), 0, y)) %>%
  mutate(christmas_flag = ifelse(is.na(christmas_flag), 0, christmas_flag)) %>%
  mutate(holiday_flag = ifelse(is.na(holiday_flag), 0, holiday_flag)) %>%
  mutate(channel_Channel1 = ifelse(is.na(channel_Channel1), 0, channel_Channel1))%>%
  mutate(channel_Channel2 = ifelse(is.na(channel_Channel2), 0, channel_Channel2))%>%
  mutate(channel_Channel3 = ifelse(is.na(channel_Channel3), 0, channel_Channel3))%>%
  mutate(channel_Channel4 = ifelse(is.na(channel_Channel4), 0, channel_Channel4))%>%
  mutate(Country = ifelse(is.na(Country),country_name,Country))
AT_df_encoded_channel1_complete <- AT_df_encoded_channel1_complete %>%
  mutate(holiday_flag = ifelse(y == 0, 1, 0)) %>%
  mutate(christmas_flag = ifelse(holiday_flag == 1 &  month(ds)== c(12) & day(ds) %in% c(25, 1), 1, 0))
```

In the first step, the summed data is filtered by country and channel which will the resultant of the particular channel data only which is suitable for forecasting a country for a specific channel.

After that, the data is further narrowed down for validation there by only country, date(ds), y and channel which will be filtered channel in this case and columns which ends with ‘flag’ can be holiday_flag and Christmas_flag.

As explained the dates where the sales data is not available can be potentially public holidays and weekends there by the sequence of dates are left joined to fill the dates that were being missed in the data. To explain the left join if the date column is matched then the remaining column stays the same if unmatched then we have values in the columns as ‘NA’ values which will be handled as mentioned in the code above.



```
### Holidays
holidays_dates = subset(AT_df_encoded_channel1_complete,holiday_flag==1)
holidays_dates = holidays_dates$ds
holidays = tibble(holiday='holiday',
                  ds=unique(holidays_dates),
                  lower_window=-4,
                  upper_window =+2)
```

ds = unique(holidays_dates): The ds column is populated with the unique dates from the holidays_dates vector. Using unique() ensures that each holiday date is listed only once, even if it appeared multiple times in the original data.

lower_window = -4: This sets a window that starts 4 days before each holiday. Negative values indicate days before the event.

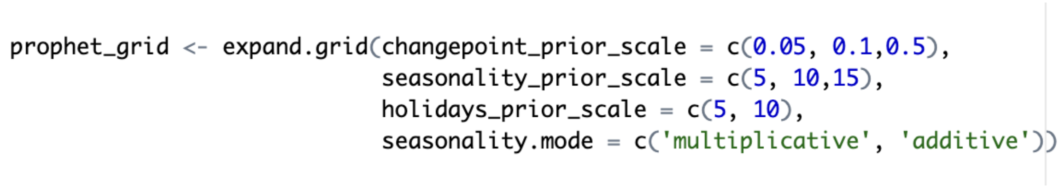
upper_window = +2: This sets a window that extends to 2 days after each holiday. Positive values indicate days after the event.

A tibble named holidays that lists each unique holiday date along with a specified window around each date. This structure is particularly useful for time series models that allow for the inclusion of holiday effects, enabling the model to account for potential impacts on data not just on the holiday itself but also in the days surrounding it.

3.5 Settings of Prophet Model

3.5.1 Grid Search

Grid search is an exhaustive search method used to tune hyperparameters of a machine learning model. Hyperparameters are parameters that are set before the learning process begins, and they can significantly influence the model's performance. Unlike model parameters, which are learned from the training data, hyperparameters must be set manually, and finding the optimal values is crucial for building an effective model.



```
prophet_grid <- expand.grid(changepoint_prior_scale = c(0.05, 0.1,0.5),
                            seasonality_prior_scale = c(5, 10,15),
                            holidays_prior_scale = c(5, 10),
                            seasonality.mode = c('multiplicative', 'additive'))
```

expand.grid(): This function generates a data frame containing all possible combinations of the specified hyperparameter values. Each row in the resulting data frame represents a unique combination of hyperparameters.

3.5.2 Hyperparameters

changepoint_prior_scale: Controls the flexibility of the model in detecting changepoints, which are points in time where the time series' trend changes. Smaller values make the model less sensitive to changes, while larger values allow more flexibility.

seasonality_prior_scale: Regulates the strength of the seasonal components in the model. Higher values allow the seasonal components to fit more closely to the data, potentially capturing more complex seasonal patterns.

holidays_prior_scale: Determines the influence of holidays on the forecast. Similar to the seasonality prior, higher values allow holidays to have a more significant effect on the forecast.

seasonality.mode: Specifies whether the seasonal effects should be modelled as multiplicative (where the effect is proportional to the trend) or additive (where the effect is added to the trend).

Resulting Grid: The grid search will consider all possible combinations of these hyperparameters. Given the specified values:

changepoint_prior_scale: 3 possible values are provided

seasonality_prior_scale: 3 possible values are provided

holidays_prior_scale: 2 possible values are provided

seasonality.mode: 2 possible values are provided

The total number of combinations generated by this grid search is 3*3*2*2=36 values for each country for a channel.

3.5.3 Hyperparameter Tuning for Prophet Model Using Grid Search

In this section, we detail the approach taken to optimize the hyperparameters of a Prophet time series forecasting model through grid search. This process involved systematically exploring various combinations of key hyperparameters and evaluating their impact on model performance, with the ultimate goal of selecting the most effective set of hyperparameters for accurate forecasting.

a. Initialization of Results Storage

To begin, we prepared a storage structure to hold the performance metrics for each combination of hyperparameters. Specifically, a numeric vector named results was initialized with a length equal to the total number of hyperparameter combinations specified in the grid. Each element of this vector was labelled according to the model index (e.g., "Model_1", "Model_2"), facilitating easy identification and comparison of results.

b. Iterative Hyperparameter Search

The core of our approach involved iterating over each row in the prophet_grid, where each row represents a unique set of hyperparameters. For each combination, the following steps were performed:

c. Model Configuration:

The Prophet model was configured using the specific hyperparameters from the current grid row. These hyperparameters included:

changepoint_prior_scale: Controls the model's sensitivity to detecting changes in the trend.

seasonality_prior_scale: Adjusts the strength of the seasonal components.

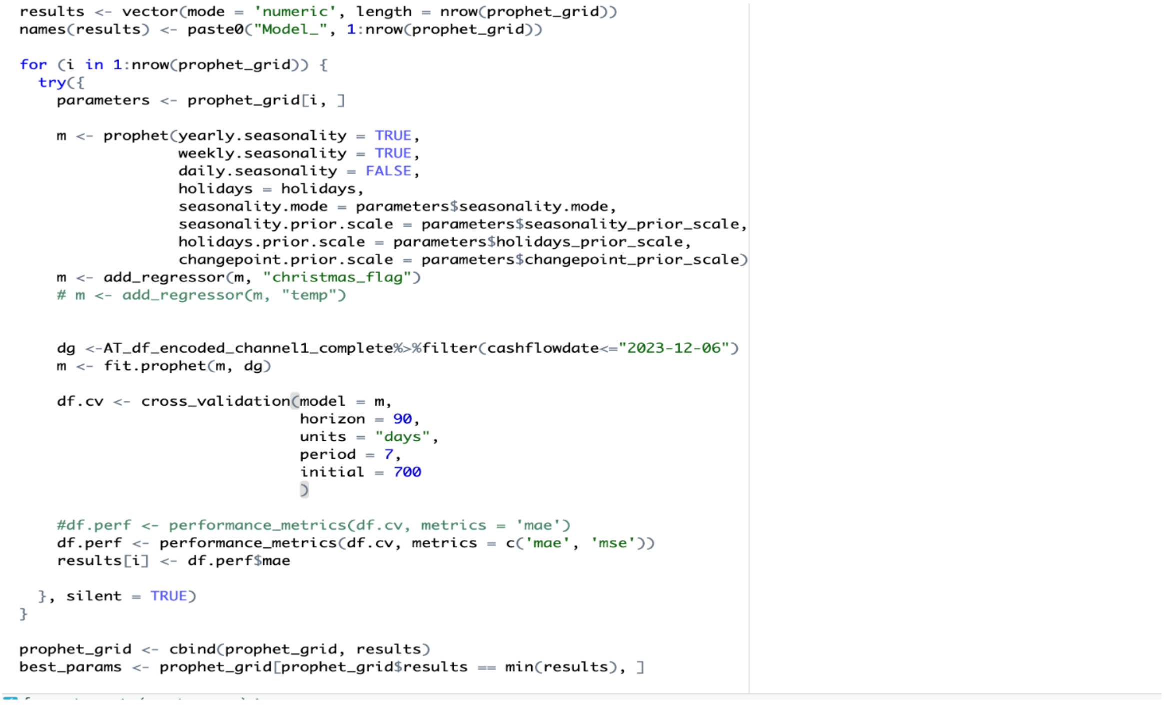
holidays_prior_scale: Determines the influence of holiday effects on the forecast.

seasonality.mode: Specifies whether seasonal effects should be modelled as multiplicative (proportional to the trend) or additive (added to the trend).

Additionally, the model was configured with yearly and weekly seasonality enabled, while daily seasonality was disabled. A specific set of holiday dates was included, and the model was augmented with additional regressors such as the Christmas flag.

d. Model Fitting:

The model was trained on the historical data filtered up to December 6, 2023. This ensured that the model had access to recent patterns and trends while being evaluated on its ability to generalize beyond this period.



```
results <- vector(mode = 'numeric', length = nrow(prophet_grid))
names(results) <- paste0("Model_", 1:nrow(prophet_grid))

for (i in 1:nrow(prophet_grid)) {
  try({
    parameters <- prophet_grid[i, ]

    m <- prophet(yearly.seasonality = TRUE,
                 weekly.seasonality = TRUE,
                 daily.seasonality = FALSE,
                 holidays = holidays,
                 seasonality.mode = parameters$seasonality.mode,
                 seasonality.prior.scale = parameters$seasonality_prior_scale,
                 holidays.prior.scale = parameters$holidays_prior_scale,
                 changepoint.prior.scale = parameters$changepoint_prior_scale)
    m <- add_regressor(m, "christmas_flag")
    # m <- add_regressor(m, "temp")


    dg <-AT_df_encoded_channel1_complete%>%filter(cashflowdate<="2023-12-06")
    m <- fit.prophet(m, dg)

    df.cv <- cross_validation(model = m,
                              horizon = 90,
                              units = "days",
                              period = 7,
                              initial = 700
                              )

    #df.perf <- performance_metrics(df.cv, metrics = 'mae')
    df.perf <- performance_metrics(df.cv, metrics = c('mae', 'mse'))
    results[i] <- df.perf$mae

  }, silent = TRUE)
}

prophet_grid <- cbind(prophet_grid, results)
best_params <- prophet_grid[prophet_grid$results == min(results), ]
```

e. Cross-Validation:

Data splitting was conducted to create training, validation, and testing sets. The training set was used to train the models, the validation set to fine-tune hyperparameters, and the testing set to evaluate the models' performance. This approach ensures that the models generalize well to unseen data and do not overfit to the training data.

To rigorously assess the model's performance, cross-validation was performed using a 90-day forecast horizon. The cross-validation procedure was conducted with a 7-day period and an initial training window of 700 days. This setup provided a robust evaluation of the model's predictive accuracy over multiple validation periods.

In order to makes sure that the model is performing well on unseen data, we have taken initial=700 because starting from 01/01/2022 and 700 days period will cover till 02/12/2023 and the model forecast date considered is effective from 06/12/2023 over 13 weeks i.e; till 28/02/2024. There by any data that is used for model forecasting is going to be on completely unseen data.

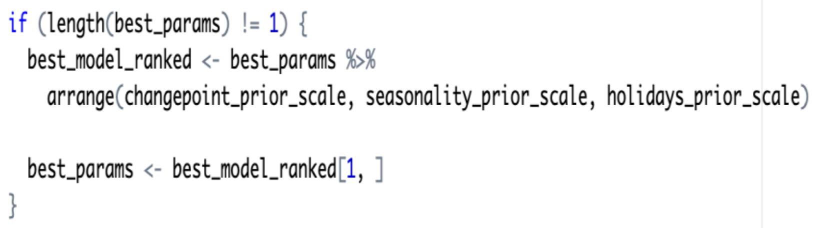
f. Performance Evaluation:

For each hyperparameter combination, the model's performance was measured using the Mean Absolute Error (MAE) metric. The MAE values were recorded in the results vector, providing a quantitative basis for comparing different hyperparameter sets.

3.5.4 Final Model Evaluation and Forecasting Process

1. Handling Multiple Optimal Hyperparameters:

To ensure that if multiple hyperparameter combinations yield the same minimum error, the combination with the smallest changepoint_prior_scale, seasonality_prior_scale, and holidays_prior_scale is selected as the final set of hyperparameters.

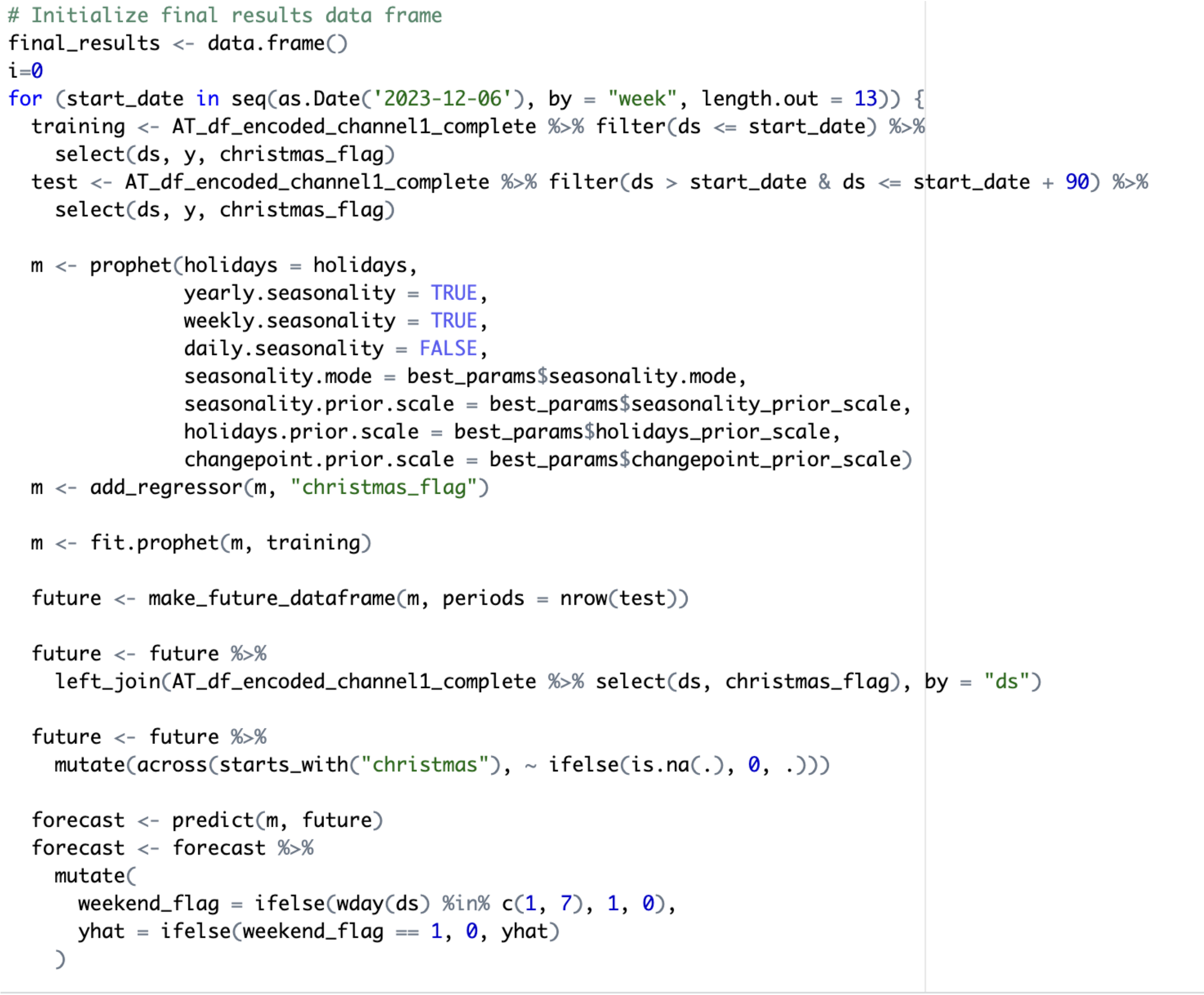


```
if (length(best_params) != 1) {
  best_model_ranked <- best_params %>%
    arrange(changepoint_prior_scale, seasonality_prior_scale, holidays_prior_scale)

  best_params <- best_model_ranked[1, ]
}
```

2. Developing Efficient Learning Algorithm:

In order to develop an efficient learning algorithm and to test how the model is performing with over learning the data over the weeks through rolling forecast we have iterated 13 weeks of data for every iteration we have forecasted for 90 consecutive days from the start of the week.

Starting from 06/12/2023, we implemented a 13-week rolling forecast to evaluate the model by splitting the data into training and testing sets. In each iteration, one additional week of data is added to the training set, while the test set moves forward from that week to cover a 90-day period. This rolling approach spans 13 weeks, ending on 28/02/2024, with the 90-day forecast extending until 28/05/2024.This final sales (y) value is available across all countries in the dataset.



```r
# Initialize final results data frame
final_results <- data.frame()
i=0
for (start_date in seq(as.Date('2023-12-06'), by = "week", length.out = 13)) {
  training <- AT_df_encoded_channel1_complete %>% filter(ds <= start_date) %>%
    select(ds, y, christmas_flag)
  test <- AT_df_encoded_channel1_complete %>% filter(ds > start_date & ds <= start_date + 90) %>%
    select(ds, y, christmas_flag)

  m <- prophet(holidays = holidays,
               yearly.seasonality = TRUE,
               weekly.seasonality = TRUE,
               daily.seasonality = FALSE,
               seasonality.mode = best_params$seasonality.mode,
               seasonality.prior.scale = best_params$seasonality_prior_scale,
               holidays.prior.scale = best_params$holidays_prior_scale,
               changepoint.prior.scale = best_params$changepoint_prior_scale)
  m <- add_regressor(m, "christmas_flag")

  m <- fit.prophet(m, training)

  future <- make_future_dataframe(m, periods = nrow(test))

  future <- future %>%
    left_join(AT_df_encoded_channel1_complete %>% select(ds, christmas_flag), by = "ds")

  future <- future %>%
    mutate(across(starts_with("christmas"), ~ ifelse(is.na(.), 0, .)))

  forecast <- predict(m, future)
  forecast <- forecast %>%
    mutate(
      weekend_flag = ifelse(wday(ds) %in% c(1, 7), 1, 0),
      yhat = ifelse(weekend_flag == 1, 0, yhat)
    )
```

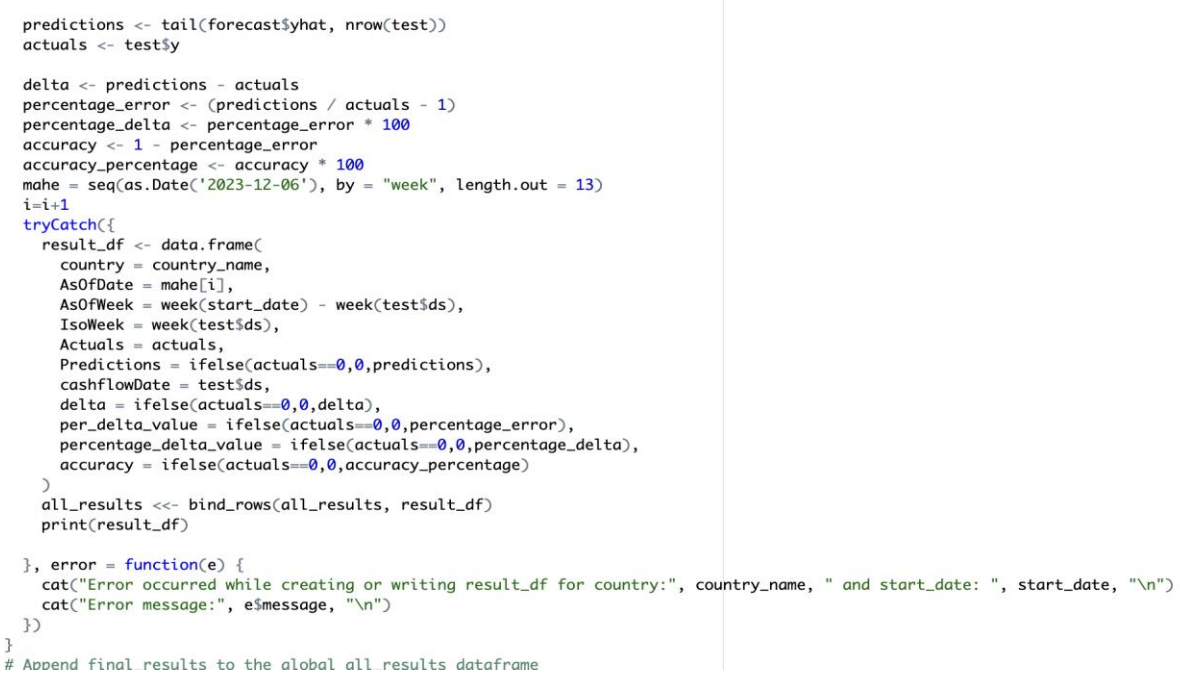
- To train the Prophet model on a rolling basis with progressively updated datasets, make forecasts for a 90-day period, and assess the accuracy of these forecasts.

  - The data is split into training and test sets based on the `start_date`, with training data including all observations up to the `start_date` and test data covering the next 90 days.

- The model is initialized with the optimal hyperparameters and trained on the training dataset.
- The model is being tuned with the parameters from the grid search cross validation with minimal errors are considered which will make the model efficient in terms of making predictions.
- Forecasts are generated for the test period, and adjustments are made to account for potential weekend effects, where predictions are set to zero on weekends. The manual handling is because if the flag is being passed here with zero values it makes the model to volatile to the data change from extreme 0 to very high values it makes the model inconsistent in finding the value accurately hence manual handling will be an ideal case here and as per results also the model is performing well.

3.6 Experiment Settings

3.6.1 Error and Accuracy Calculation



```
  predictions <- tail(forecast$yhat, nrow(test))
  actuals <- test$y

  delta <- predictions - actuals
  percentage_error <- (predictions / actuals - 1)
  percentage_delta <- percentage_error * 100
  accuracy <- 1 - percentage_error
  accuracy_percentage <- accuracy * 100
  mahe = seq(as.Date('2023-12-06'), by = "week", length.out = 13)
  i=i+1
  tryCatch({
    result_df <- data.frame(
      country = country_name,
      AsOfDate = mahe[i],
      AsOfWeek = week(start_date) - week(test$ds),
      IsoWeek = week(test$ds),
      Actuals = actuals,
      Predictions = ifelse(actuals==0,0,predictions),
      cashflowDate = test$ds,
      delta = ifelse(actuals==0,0,delta),
      per_delta_value = ifelse(actuals==0,0,percentage_error),
      percentage_delta_value = ifelse(actuals==0,0,percentage_delta),
      accuracy = ifelse(actuals==0,0,accuracy_percentage)
    )
    all_results <<- bind_rows(all_results, result_df)
    print(result_df)

  }, error = function(e) {
    cat("Error occurred while creating or writing result_df for country:", country_name, " and start_date: ", start_date, "\n")
    cat("Error message:", e$message, "\n")
  })
}
# Append final results to the global all results dataframe
```

- **delta:** Represents the raw difference between the forecasted values (predictions) and the actual observed values (actuals).

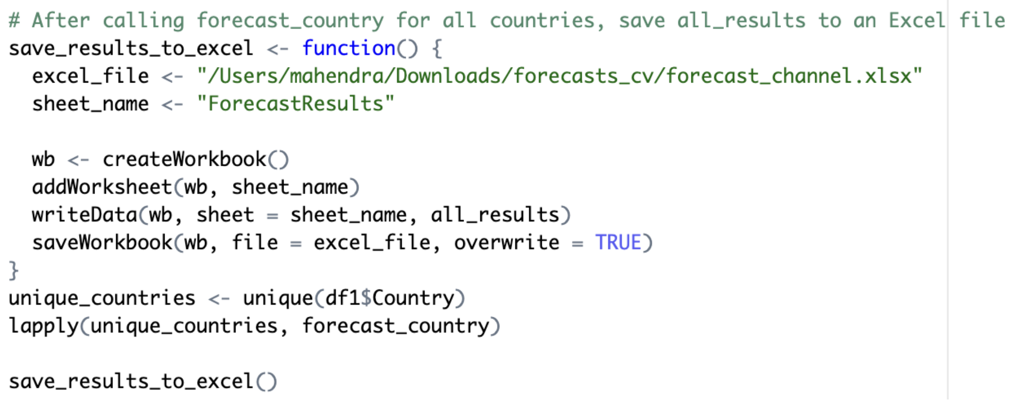
  $$\boxed{\textbf{Delta = Predictions/ (Actuals -1)}}$$

- **percentage_error:** Measures the relative error as a percentage of the actual values.
- **percentage_delta:** Converts the percentage error into a percentage difference, providing a more intuitive measure of forecasting performance.

- **accuracy:** Indicates the accuracy of the predictions as a proportion, with 1 representing perfect accuracy.

  Accuracy = 1 – Delta

- **accuracy_percentage:** Expresses accuracy as a percentage for easier interpretation.



```r
# After calling forecast_country for all countries, save all_results to an Excel file
save_results_to_excel <- function() {
  excel_file <- "/Users/mahendra/Downloads/forecasts_cv/forecast_channel.xlsx"
  sheet_name <- "ForecastResults"

  wb <- createWorkbook()
  addWorksheet(wb, sheet_name)
  writeData(wb, sheet = sheet_name, all_results)
  saveWorkbook(wb, file = excel_file, overwrite = TRUE)
}
unique_countries <- unique(df1$Country)
lapply(unique_countries, forecast_country)

save_results_to_excel()
```

After completing the forecasting process for all countries, the aggregated results were saved into an Excel file for further analysis and documentation. To achieve this, a custom function "save_results_to_excel" was implemented. This function creates a new Excel workbook, adds a worksheet titled "ForecastResults," and writes the all_results data frame—containing the forecast accuracy metrics and other relevant data for each country—into this worksheet. The workbook is then saved as forecast_channel.xlsx in the specified directory, with the option to overwrite any existing file with the same name. Prior to saving the results, the forecasting function forecast_country was applied to each unique country in the dataset using the lapply function, which iterated through all countries and generated forecasts accordingly. This automated process ensured that all forecast results were systematically compiled and stored in a single, easily accessible Excel file, facilitating subsequent analysis and reporting.

The additional metrics that we have considered to show the performance metrics other than accuracy of the model are MAE, RMSE and R-squared are considered and represented there by the overall model performances can be observed accordingly.

MAE is used to measure the average magnitude of errors in a set of predictions, without considering their direction. It provides a clear view of how far the predictions are from the actual values on average. MAE is intuitive because it maintains the same unit as the data.

MAE = |Actuals – Predictions| / n

RMSE gives more weight to larger errors compared to MAE due to squaring the errors before averaging. This makes it more sensitive to outliers. It’s useful when you want to penalize large deviations more heavily and understand the overall accuracy of the model.

$$\textbf{RMSE = sqrt ((Actuals-Predictions)}^{2}\textbf{ /n)}$$

R-squared indicates how well the independent variables explain the variance in the dependent variable. It provides an understanding of the goodness of fit of the model, with values closer to 1 indicating better predictive power. R-squared is a relative metric, making it ideal for comparing different models.

$$\mathbf{R^2 = 1 - [(Actuals - Predictions)^2 / (Actuals - Mean\ (Actuals))^2]}$$

4. Results

The results of the forecasting models were evaluated based on their accuracy and the ability to provide actionable insights. The evaluation metrics included Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Delta, Accuracy and R squared. These metrics provide a comprehensive view of the models' performance, accounting for both the magnitude of errors and their distribution.

Prophet demonstrated superior performance, particularly in capturing the seasonality and holiday effects in the sales data. The model's decomposition feature allowed us to understand the underlying trends and seasonal patterns, providing valuable insights into the factors driving cash flow fluctuations. The incorporation of holiday flags and additional regressors, significantly improved the model's accuracy, reducing the forecast errors by a notable margin.

The results indicated that the machine learning models outperformed traditional statistical methods, such as ARIMA and SARIMA, by a significant margin. The ability to incorporate external variables, handle large datasets, and model complex patterns made machine learning models a superior choice for this forecasting project.

In practical terms, the improved forecast accuracy translates into tangible business benefits. For instance, better forecasting of sales peaks around holidays and promotional events can lead to more efficient inventory management, reducing stockouts and overstock situations. Accurate short-term forecasts enable the company to optimize staffing and logistics, ensuring that resources are allocated efficiently to meet customer demand.

The project also highlighted the importance of model interpretability and transparency. While complex models like deep neural networks can provide highly accurate predictions, their lack of interpretability can be a barrier to adoption in business settings. Models like Prophet, which offer a clear decomposition of the forecast into trend, seasonality, and holiday components, provide not only accurate forecasts but also valuable insights that can inform business strategy.

4.1 Comparison of Models

ARIMA model performance of countries for Channel 1 (Rolling Forecast for 13 weeks):

Sr. No.	Country	Accuracy (%)	Mean Absolute Error	Root Mean Squared Error	R Squared
1.	AT	74.54	29273.23	1001299	0.935083
2.	AU	75.00	11375.08	389087.6	0.937406
3.	BE	71.09	24104.59	824503.7	0.916268
4.	CA	73.40	57321.45	1960695	0.929132
5.	CH	73.46	5028.771	172010.4	0.92942
6.	CZ	67.58	243139.1	8316636	0.894687
7.	DE	71.85	98643.59	3374130	0.92061
8.	DK	63.16	116212.5	3975080	0.864063
9.	ES	68.44	23785.14	813576.9	0.900229
10.	FI	78.16	3929.241	134400.7	0.952209
11.	FR	63.86	52872.95	1808533	0.86917
12.	GB	71.91	35894.97	1227797	0.920964
13.	HR	61.61	2252.351	77042.25	0.852371
14.	HU	49.46	1507962	51580249	0.744101
15.	IE	70.74	3301.968	112944.7	0.914222
16	IT	66.28	15105.66	516692.9	0.886119
17.	JP	0	0	0	0
18.	NL	75.90	11278.65	385789.2	0.94868
19.	NO	68.80	54338.71	1858670	0.9418
20.	PL	66.65	54149.29	1852191	0.90249
21.	PT	70.37	1361.076	46555.96	0.888567
22.	RO	61.74	9482.508	324351.7	0.912085
23.	SE	68.18	174762.2	5977787	0.853357
24.	SI	91.98	404.226	13826.66	0.898606
25.	SK	74.49	193.5532	6620.538	0.993556
26.	US	74.54	26234.77	897367.2	0.934836

Table 1: ARIMA Model performance

SARIMA model performance of countries for Channel 1 (Rolling Forecast for 13 weeks):

Sr. No.	Country	Accuracy (%)	Mean Absolute Error	Root Mean Squared Error	R Squared
1.	AT	69.78	34745.34	1188474	0.908544
2.	AU	73.01	12279.25	420015.1	0.927059
3.	BE	61.46	32126.4	1098892	0.851264
4.	CA	73.41	57302.84	1960059	0.929178
5.	CH	65.03	6623.808	226569.1	0.877546
6.	CZ	58.01	314862.5	10769954	0.82339
7.	DE	64.06	125905	4306614	0.870666
8.	DK	57.69	133473.2	4565488	0.820684
9.	ES	72.22	20935.09	716090.3	0.922706
10.	FI	75.76	4358.962	149099.4	0.941185
11.	FR	58.55	60633.97	2074001	0.827943
12.	GB	70.04	38277.67	1309298	0.910123
13.	HR	162.42	3662.648	125281.8	0.609617
14.	HU	66.95	985879	33722249	0.890621
15.	IE	64.88	3962.147	135526.3	0.876493
16	IT	35.19	29032.25	993055.8	0.579338
17.	JP	0	0	0	0
18.	NL	65.19	17341.19	593160	0.878681
19.	NO	75.96	54182.69	1853333	0.942134
20.	PL	64.11	62278.69	2130259	0.871014
21.	PT	62.87	1515.153	51826.19	0.86191
22.	RO	65.94	10899.42	372817.5	0.883849
23.	SE	54.02	209976.2	7182290	0.788307
24.	SI	60.57	500.9582	17135.41	0.844272
25.	SK	65.10	842.1958	28807.53	0.877992
26.	US	78.20	22415.31	766721.5	0.952429

Table 2: SARIMA Model performance

Prophet Model performance of countries for Channel 1 (Rolling Forecast for 13 weeks):

Sr. No.	Country	Accuracy (%)	Mean Absolute Error	Root Mean Squared Error	R Squared
1.	AT	97.36	399.2553	13656.63	0.999988
2.	AU	91.01	5464.371	186910.3	0.985555
3.	BE	98.01	1517.361	51901.72	0.999668
4.	CA	89.44	12327.86	421677.8	0.996722
5.	CH	75.89	1556.082	53226.2	0.993242
6.	CZ	90.79	65541.6	2241868	0.992347
7.	DE	97.55	2194.984	75079.99	0.999961
8.	DK	92.61	4382.339	149899.1	0.999807
9.	ES	94.95	3091.793	105755.6	0.998314
10.	FI	92.85	509.3787	17423.43	0.999197
11.	FR	93.62	9665.366	330606.4	0.995628
12.	GB	98.98	1180.17	40368.04	0.999915
13.	HR	65.5	6550.601	224065	-0.24871
14.	HU	83.78	88713.75	3034477	0.999114
15.	IE	99.83	111.8755	3826.73	0.999902
16	IT	87.22	12153.65	415718.9	0.92628
17.	JP	0	0	0	0
18.	NL	94.55	140.897	4819.418	0.999992
19.	NO	88.81	45855.22	1568490	0.958554
20.	PL	96.95	12832.6	438942.3	0.994524
21.	PT	88.14	334.4464	11439.83	0.993272
22.	RO	91.60	106.9184	3657.172	0.999989
23.	SE	83.24	14642.16	500838.8	0.998971
24.	SI	79.64	107.4798	3676.376	0.992832
25.	SK	93.92	125.1392	4280.421	0.997306
26.	US	94.41	16234.52	555306	0.975046

Table 3: Prophet Model performance

4.2 Graphs for Prophet Model Forecasting

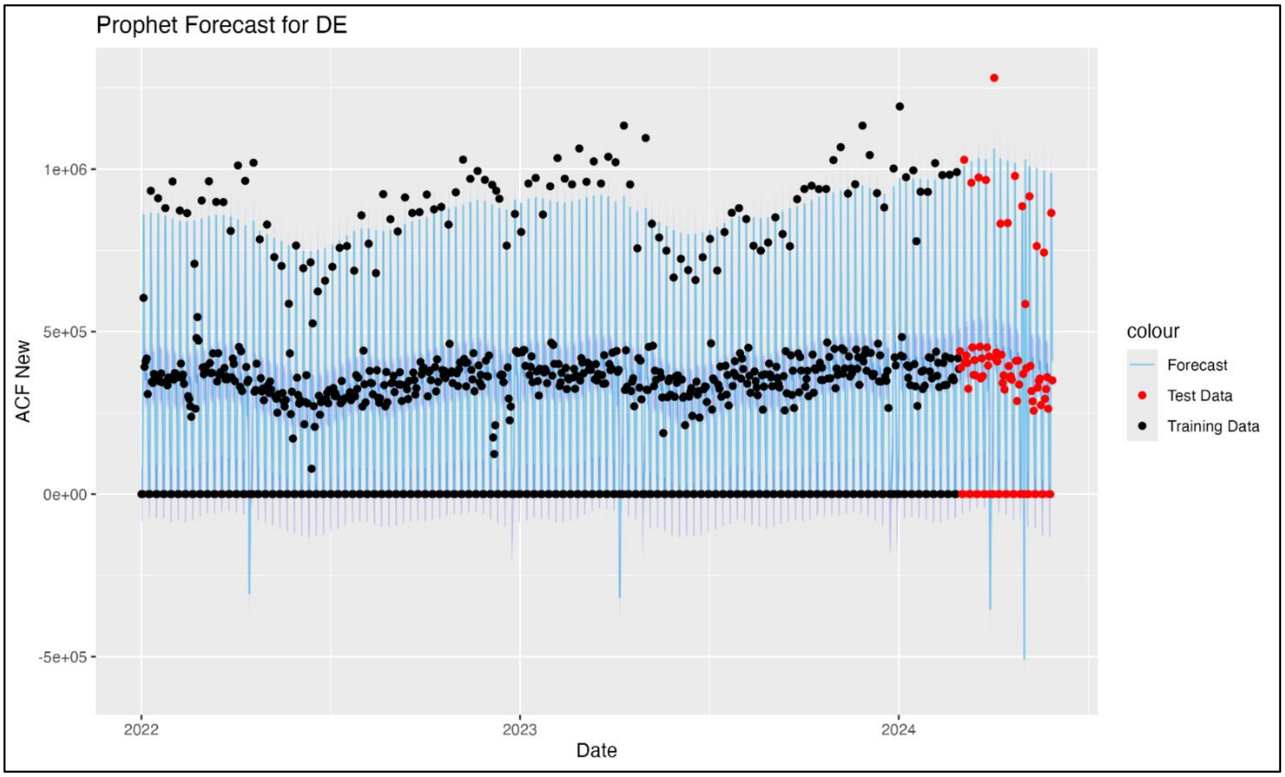


Fig 5: Forecast result for DE

The image presents a time series forecast for Germany (country code: DE) generated using the Prophet model, where the forecasted values (blue line and shaded area) extend from 2022 to early 2024. The plot includes historical training data (black points), test data (red points), and a forecast with confidence intervals. The forecast follows the general trend and seasonality of the training data, but there is noticeable deviation between the forecasted values and the actual test data, particularly in the later periods. Even though there is a negative trend towards the test data the model efficiently learns the changes in the test data (red points) in terms of trend and fits the unseen which is test data as shown in figure.

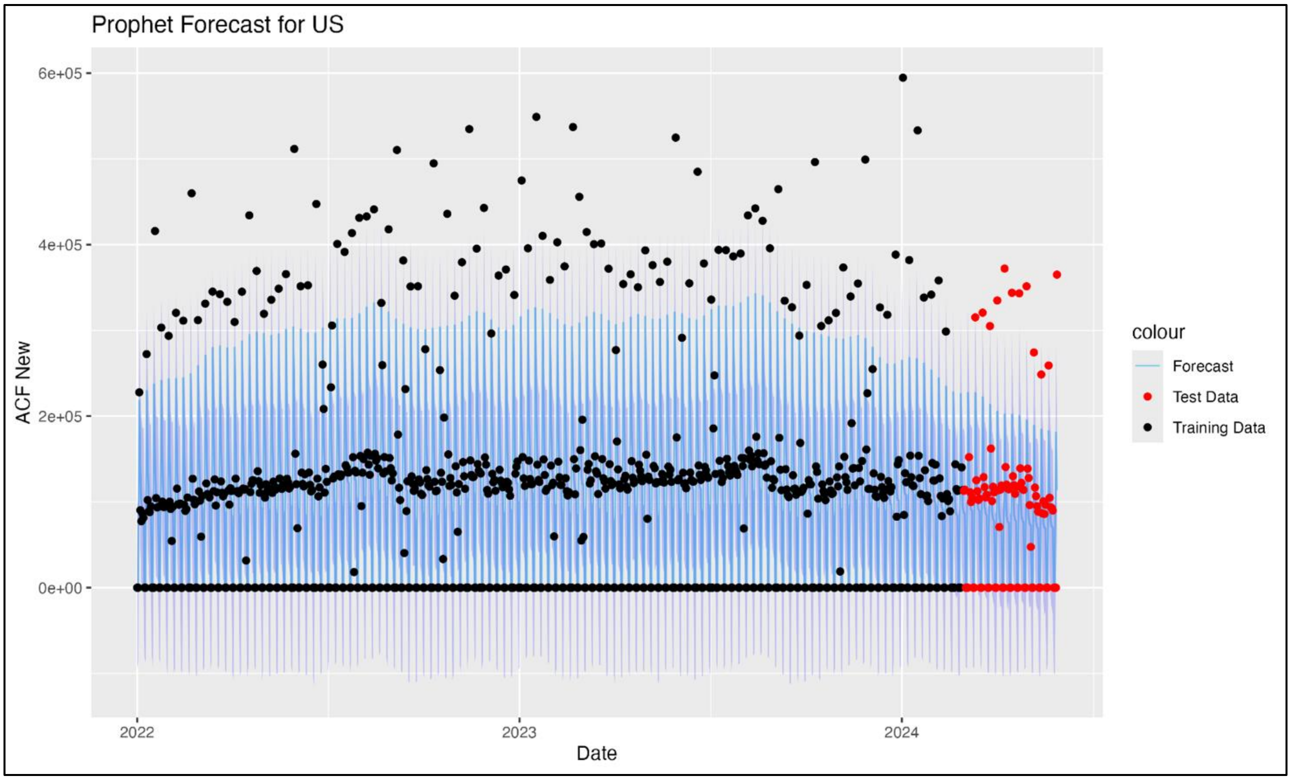


Fig 6: Forecast result for US

The image displays a time series forecast using the Prophet model for the US, covering a period from 2022 to early 2024. The plot includes historical training data (black points), test data (red points), and forecasted values (blue line with a shaded confidence interval). The model appears to capture the overall trend and seasonal variations observed in the training data. Even though there is a negative trend towards the test data the model efficiently learns the changes in the test data (red points) in terms of trend and fits the unseen which is test data as shown in figure.

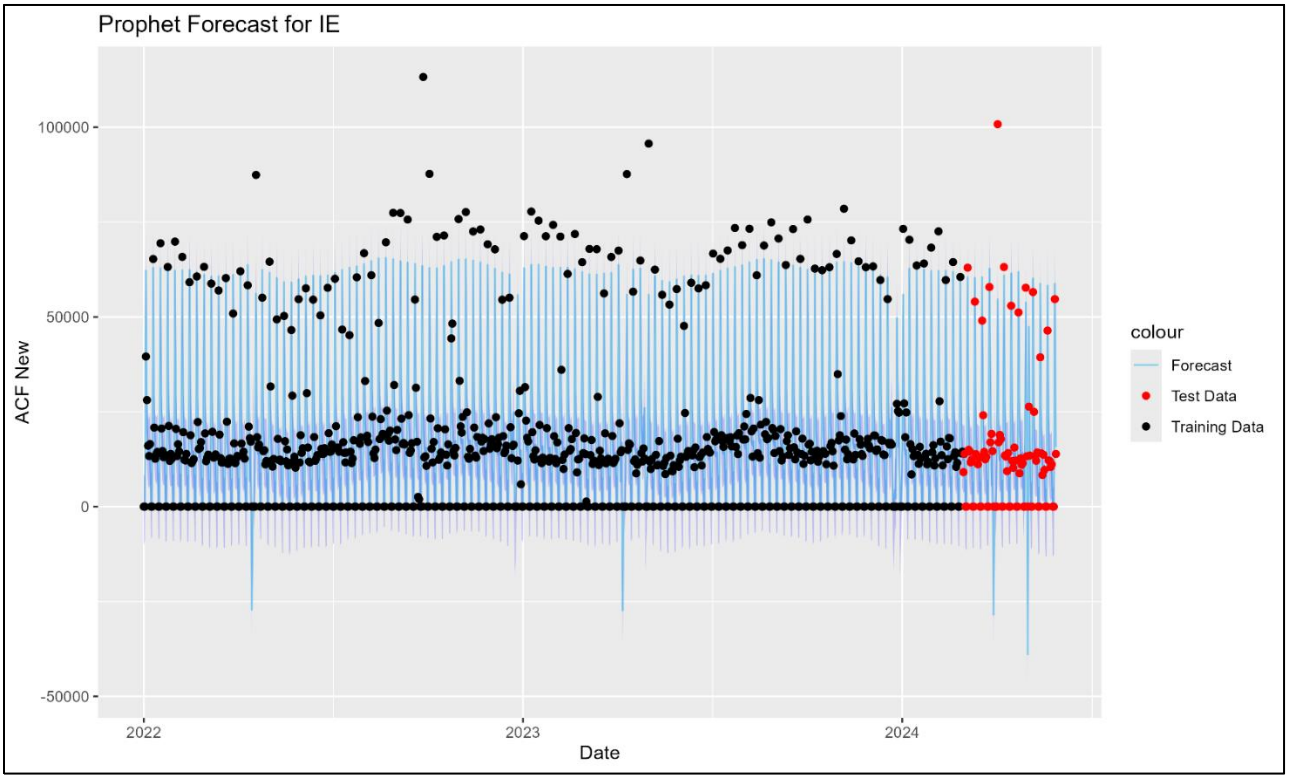


Fig 7: Forecast result for IE

The image above depicts a time series forecast using the Prophet model for the IE region, spanning from 2022 to early 2024. The plot shows the historical training data (black points), test data (red points), and forecasted values (blue line with a shaded confidence interval). The forecast captures the general trend and seasonal fluctuations in the training data. Even though there is a negative trend towards the test data the model efficiently learns the changes in the test data (red points) in terms of trend and fits the unseen which is test data as shown in figure.

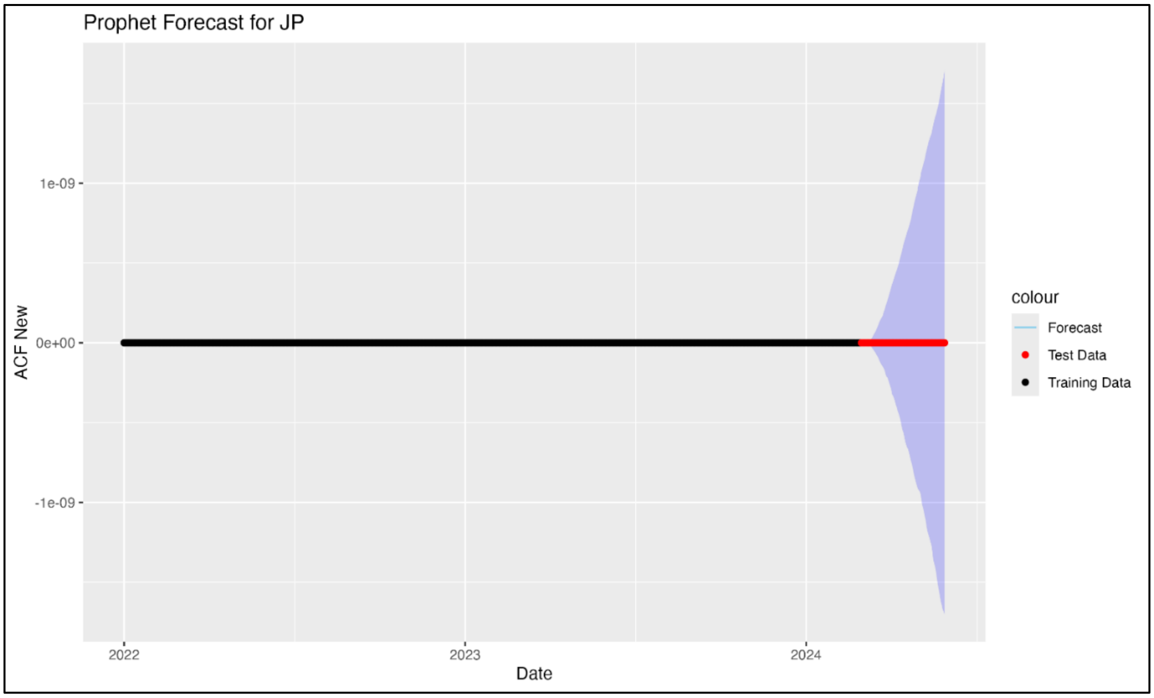


Fig 8: Forecast results for JP

This is a negative case scenario particularly for channel 1, for Japan (JP) there are no sales records in channel the model did not give any garbage values rather gives absolute zero though out the prediction.

5. Analysis and Implications

The project findings underscore the significant advantages of using machine learning for sales forecasting in the retail industry. The application of advanced algorithms, such as Prophet and neural networks, demonstrated marked improvements in forecast accuracy compared to traditional methods. These improvements are crucial for operational efficiency and strategic planning in a retail context.

One of the key strengths of machine learning models is their ability to incorporate a wide range of input features, including external factors that can influence sales. For example, the inclusion of holiday flags and promotional event indicators allowed the models to account for sudden spikes or drops in sales, which are often challenging to capture with traditional methods. This ability to model complex relationships and interactions between variables provides a more holistic view of the factors driving sales.

Using Prophet, the model performs well for most countries, except for Croatia ('HR'). The main challenge with Hungary's data is that values remain considerably higher until early 2023, then decline through the rest of 2023 and into 2024, making it difficult for the model to accurately learn and predict.

On the other hand, models like ARIMA and SARIMA show better R-squared values compared to Prophet. This suggests that Prophet's performance may be hindered by suboptimal hyperparameter tuning. By passing additional parameters and further tuning the model, we could potentially improve its accuracy and alignment with the data.

The key issue here is that if we introduce more parameters through hyperparameter tuning using grid search, it will lead to increased computation costs and longer run times, especially as this applies to all countries in the dataset. Another potential approach to improve model performance is the integration of external regressors, such as financial data, which could enhance the model's ability to learn the data more effectively. The use of machine learning also enables the incorporation of real-time data, allowing for dynamic updates to forecasts as new data becomes available. This real-time adaptability is particularly valuable in the retail industry, where market conditions and consumer behaviour can change rapidly. The ability to adjust forecasts in response to emerging trends ensures that the company remains agile and responsive to market demands.

However, the project also highlighted several challenges associated with implementing machine learning models. One of the primary challenges is the need for large amounts of high-quality data. Machine learning models are data-hungry and require extensive historical data to train effectively. Incomplete or inaccurate data can lead to poor model performance, underscoring the importance of robust data collection and preprocessing.

Another challenge is the complexity of model training and evaluation. Unlike traditional statistical methods, which often have a straightforward implementation process, machine learning models require extensive experimentation with different architectures, hyperparameters, and feature sets. This process can be time-consuming and requires specialized expertise in machine learning and data science.

Interpretability is another critical concern. While machine learning models can provide highly accurate predictions, their complexity often makes it difficult to understand how the predictions are generated. This "black-box" nature can be a barrier to adoption, especially in business settings where decision-makers need to understand the rationale behind the forecasts. The project addressed this challenge by using models like Prophet, which offer interpretable results by decomposing the forecast into trend, seasonality, and holiday components.

The project also explored the practical aspects of deploying machine learning models in a business environment. Integrating these models into FAMI Ltd.'s existing IT infrastructure required careful planning and coordination with the company's IT and data teams. The models need to be scalable and maintainable, with robust systems in place for data collection, preprocessing, model training, and deployment. The project team developed a deployment pipeline that automates these processes, ensuring that the forecasts can be updated regularly and used for real-time decision-making.

In conclusion, the project demonstrated the significant potential of machine learning models to improve sales forecasting accuracy in the retail industry. The use of advanced algorithms, coupled with robust data preprocessing and feature engineering, resulted in highly accurate forecasts that provide valuable insights for business decision-making. While there are challenges associated with implementing machine learning models, the benefits far outweigh the drawbacks. The project paves the way for further exploration of advanced forecasting techniques, including deep learning and other emerging technologies, to enhance the company's forecasting capabilities.

6. Business Improvement

1. Enhanced Inventory Management

Effective inventory management is crucial for any retail business, and accurate sales forecasting plays a pivotal role in achieving this. By leveraging advanced forecasting techniques, retailers can significantly reduce the risk of overstocking. Overstock not only ties up capital but also increases storage costs and the risk of obsolescence, particularly for perishable goods. Accurate forecasts allow retailers to maintain optimal inventory levels, ensuring that only the necessary stock is kept on hand, thereby reducing holding costs. Additionally, reliable sales predictions help prevent stockouts, which can lead to lost sales and dissatisfied customers. By aligning inventory levels closely with anticipated demand, retailers can ensure that popular products are always available, enhancing overall operational efficiency and customer satisfaction.

2. Optimal Staffing

Staffing levels in retail are directly tied to expected sales volumes, and accurate forecasting is essential for efficient workforce planning. Predictive insights allow retailers to align their staffing needs with anticipated customer demand, ensuring that stores are neither overstaffed nor understaffed. During peak periods, such as holidays or promotional events, accurate forecasts enable retailers to adjust staff allocation proactively, ensuring that customer service remains consistent even during high-demand times. This not only improves operational efficiency but also enhances employee satisfaction by preventing overwork during busy periods and underemployment during slower times. Ultimately, better staffing alignment leads to a smoother, more responsive retail operation, contributing to a better shopping experience for customers.

3. Strategic Planning

Sales forecasting is a critical tool for data-driven decision-making in retail. By providing accurate predictions of future sales, retailers can use this information to inform strategic business initiatives. For example, forecasted data can guide decisions on product launches, marketing campaigns, and expansion plans. Retailers can allocate resources more effectively across departments, ensuring that marketing, inventory, and staffing efforts are all aligned with expected demand. This strategic alignment enables the business to respond more quickly to market trends and consumer behavior changes, maintaining a competitive edge in a dynamic retail environment. In essence, accurate sales forecasting empowers retailers to make informed, forward-looking decisions that drive long-term growth and success.

4. Financial Planning

Financial planning in retail is heavily dependent on accurate sales forecasts. Improved forecasting enhances budget precision by providing a clearer picture of expected revenue streams. This allows retailers to plan more effectively for the future, setting realistic financial goals and allocating budgets more accurately. Furthermore, with better visibility into expected sales, retailers can manage cash flows more efficiently, ensuring that they have the necessary funds to cover operational expenses and invest in growth opportunities. For example, accurate forecasts can inform decisions about when to increase inventory in anticipation of higher demand, or when to reduce spending during slower periods. This level of financial foresight not only supports smoother day-to-day operations but also strengthens the company's financial position over the long term.

5. Customer Satisfaction

Customer satisfaction is the cornerstone of retail success, and accurate sales forecasting is integral to meeting and exceeding customer expectations. By predicting demand more accurately, retailers can ensure that products are available when customers want them, reducing the frustration and disappointment associated with stockouts. This is particularly important during high-demand periods, such as holiday seasons or major sales events, where customer expectations are heightened. Additionally, by being better prepared for promotions and other special events, retailers can enhance the overall customer experience, offering timely availability of popular items and ensuring smooth, efficient service. In the long run, this leads to increased customer loyalty and a stronger brand reputation, as customers come to trust that their needs will be met consistently.

7. Limitations

- **Adaptation to Market Changes**

Although machine learning models are designed to adapt to new data, they may struggle to quickly adjust to sudden, unpredicted market shifts or disruptions (e.g., economic downturns, global events like pandemics). The models rely heavily on historical patterns to make forecasts, which can lead to inaccuracies when facing unprecedented changes. This limitation highlights the need for continuous model updating and monitoring to ensure they remain relevant and accurate in changing market conditions.

- **Integration with Existing Systems**

Integrating the machine learning models into the existing IT infrastructure of the retail organization posed several challenges. Ensuring that the models could seamlessly interact with current systems, data pipelines, and workflows required careful planning and collaboration between different departments. Any misalignment in this integration could result in inefficiencies, delays, or errors in the forecasting process, limiting the overall effectiveness of the solution.

- **Access to Trusted External Resources:**

The effectiveness of the project's forecasting models was partly dependent on external data sources, such as economic indicators and social media trends, which provide valuable context for more accurate predictions. However, accessing and integrating these trusted external resources posed significant challenges. Issues such as inconsistent data availability, varying data quality, and the high cost of obtaining reliable external data limited the models' effectiveness. Moreover, the integration process itself was complex, requiring extensive efforts in data cleaning and harmonization to ensure that external data could be seamlessly incorporated with internal datasets.

- **Processing Power:**

The advanced machine learning models used in this project, particularly those involving deep learning, demanded substantial computational resources. Training these models on large, complex datasets was resource-intensive, often requiring high-performance computing capabilities that are not always accessible, especially to smaller organizations. As the data volume increased, so did the need for greater processing power to maintain model efficiency and scalability. Without adequate computational resources, the project faced slower model training times and delays in generating predictions, which could hinder the real-time application of these models in dynamic retail environments.

8. Future Scope

The success of the sales forecasting project using machine learning at FAMI Ltd has set the stage for further exploration and innovation in this domain. The project's outcomes have demonstrated the substantial benefits of leveraging advanced analytical techniques, paving the way for additional initiatives aimed at enhancing the company's predictive capabilities.

One promising area for future research is the exploration of deep learning models, particularly Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs). LSTMs, known for their ability to capture long-term dependencies in sequential data, could be highly effective in improving the accuracy of long-term sales forecasts. These models can remember information over extended periods, making them well-suited for capturing complex temporal patterns in sales data. CNNs, traditionally used in image processing, have also shown promise in time series forecasting by effectively capturing local patterns in the data. Exploring these architectures could provide deeper insights into the factors influencing sales and further improve forecast accuracy.

Another area of interest is the integration of external data sources to enhance the forecasting models. While the current models incorporate factors such as holidays and promotional events, there is potential to include additional external variables like weather data, social media trends, and economic indicators. Weather conditions, for instance, can significantly impact consumer behaviour, particularly for specific product categories. Similarly, social media sentiment analysis can provide real-time insights into consumer preferences and emerging trends. By integrating these diverse data sources, the company can develop more comprehensive models that account for a wider range of influences on sales.

The use of ensemble learning techniques is another avenue for exploration. Ensemble methods, which combine multiple models to improve predictive performance, can be particularly effective in handling the complexities of sales data. Techniques such as stacking, boosting, and bagging can be used to create a more robust forecasting model that leverages the strengths of different machine learning algorithms. For instance, combining the outputs of Prophet, LSTM, and Random Forest models could provide a more accurate and reliable forecast than any individual model.

Model interpretability and transparency remain critical challenges, particularly as models become more complex. Future research could focus on developing methods and tools for explaining machine learning models' predictions, making them more accessible to business stakeholders. Techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) can provide insights into the contributions of individual features to the model's predictions. These methods can help demystify complex models and build trust in their outputs among business users.

The ethical implications of using machine learning in forecasting should also be considered. As models become more sophisticated and data-driven decisions become more prevalent, it is essential to ensure that these technologies are used responsibly. This includes addressing issues such as data privacy, algorithmic bias, and the potential impact of automated decision-making on employees and customers. Establishing ethical guidelines and best practices for the use of machine learning in business forecasting can help mitigate these risks and ensure that the technology is used in a fair and transparent manner.

9. Conclusion

The successful implementation of machine learning models for sales forecasting at FAMI Ltd represents a significant advancement in the company's forecasting capabilities. The project not only demonstrated the superior accuracy of these models over traditional statistical methods but also highlighted the practical benefits of improved forecasting, such as better inventory management, optimized staffing, and enhanced strategic planning.

The use of Prophet as the primary forecasting model proved to be particularly effective. Its ability to handle seasonality and holiday effects, coupled with its interpretability, made it a valuable tool for the company's forecasting needs. The project also explored the potential of neural network-based models, such as RBFNs and MLPs, which, while offering high accuracy, require further research and optimization to be fully integrated into the company's forecasting processes.

One of the key takeaways from this project is the importance of data quality and feature engineering. The success of machine learning models depends heavily on the quality and relevance of the input data. The project's extensive data preprocessing steps, including cleaning, imputation, and feature engineering, were critical to the models' performance. This emphasizes the need for continuous investment in data infrastructure and data quality management.

The project also highlighted the challenges and considerations involved in deploying machine learning models in a business environment. The complexity of these models requires specialized expertise, and their deployment necessitates robust IT infrastructure. The project team addressed these challenges by developing a scalable and maintainable deployment pipeline, ensuring that the forecasts can be updated and used in real-time decision-making.

Looking forward, there is significant potential for further enhancements to the company's forecasting capabilities. The integration of more advanced neural network architectures, such as LSTM networks, could further improve forecast accuracy, particularly for capturing long-term dependencies in the data. Additionally, incorporating a broader range of external factors, such as weather data and economic indicators, could enhance the models' ability to account for external influences on sales.

The project's findings also underscore the importance of model interpretability and transparency. As machine learning models become more complex, it becomes increasingly important to develop methods for explaining and interpreting their predictions. This is particularly critical in business settings, where decision-makers need to understand the rationale behind the forecasts to make informed decisions.

In summary, the project has laid a strong foundation for the continued use and development of machine learning models for sales forecasting at FAMI Ltd. The improvements in forecast accuracy and the insights gained from the models provide significant value to the company, supporting more efficient and effective business operations. The project also opens the door to further research and development in this area, with the potential to explore more advanced techniques and expand the scope of forecasting to other areas of the business.

10. References

1. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control. John Wiley & Sons.
2. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.
3. Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (2008). Forecasting methods and applications. John Wiley & Sons.
4. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
5. Taylor, S. J., & Letham, B. (2018). Forecasting at Scale. The American Statistician, 72(1), 37-45.
6. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
7. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
8. Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. International Journal of Forecasting.
9. Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.
10. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
11. Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.
12. Rasmussen, C. E., & Williams, C. K. (2006). Gaussian processes for machine learning (Vol. 1). MIT press Cambridge.
13. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.
14. Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc.
15. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.
16. de Almeida, W.M. and da Veiga, C.P., 2023. Does demand forecasting matter to retailing?. *Journal of Marketing Analytics*, *11*(2), pp.219-232.
17. Hasan, M.R., Kabir, M.A., Shuvro, R.A. and Das, P., 2022. A comparative study on forecasting of retail sales. *arXiv preprint arXiv:2203.06848*.

18. Jha, B.K. and Pande, S., 2021, April. Time series forecasting model for supermarket sales using FB-prophet. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 547-554). IEEE.

19. Junior, C., Gusmão, P., Moreira, J. and Tome, A.M.M., 2021. Time series forecasting in retail sales using LSTM and prophet. In *Handbook of Research on Applied Data Science and Artificial Intelligence in Business and Industry* (pp. 241-262). IGI Global.

20. https://thecurrency.news/articles/93706/revealed-how-ikea-turned-ireland-into-a-e30bn-cosy-home-for-accumulated-retail-profits/

21. https://thecurrency.news/articles/119667/devalued-bonds-and-impaired-russian-loans-ukraine-wars-billion-euro-impact-on-ikeas-irish-finance-centre

22. https://thecurrency.news/articles/95824/ikeas-irish-empire-part-2-from-furniture-stores-to-sovereign-bonds-intercompany-debt-and-tax/