SALES FORECASTING MODELS FOR DIRECT SELLING BUSINESS: A DATA-DRIVEN APPROACH TO PREDICTIVE ANALYTICS

SIVARAJAN A/L S.ESVARAN

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name	:	SIVARAJAN A/L S.ESVARAN			
Student's Matric No.	:	MCS241051	Academic Session	:	20242025-02
Date of Birth	:	17TH JANUARY 1993	UTM Email	:	sivarajan@graduate.utm.my
Thesis Title	:	SALES FORECASTING MODELS FOR DIRECT SELLING BUSINESS: A DATA-DRIVEN APPROACH TO PREDICTIVE ANALYTICS			

I declare that this thesis is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this thesis for academic exchange.

Signature of Student:

Signature :

Full Name : SIVARAJAN A/L S.ESVARAN
Date : 30th JUNE 2025

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I	Full Name of Supervisor II
_	_

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

Date:

Librarian

Jabatan Perpustakaan UTM,

Universiti Teknologi Malaysia,

Johor Bahru, Johor

Sir,

CLASSIFICATION OF THESIS AS RESTRICTED/CONFIDENTIAL

TITLE: SALES FORECASTING MODELS FOR DIRECT SELLING BUSINESS:
A DATA-DRIVEN APPROACH TO PREDICTIVE ANALYTICS

AUTHOR'S FULL NAME:SIVARAJAN A/L S.ESVARAN

Please be informed that the above-mentioned thesis titled Sales Forecasting Models For Direct Selling Business: A Data-Driven Approach To Predictive Analytics should be classified as RESTRICTED/CONFIDENTIAL for a period of three (3) years from the date of this letter. The reasons for this classification are

(i)

(ii)

(iii)

Thank you.

Yours sincerely,

SIGNATURE:

NAME:

ADDRESS OF SUPERVISOR:

"I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in term of scope and quality for the award of the degree of Master in (Data Science)"

Signature : _____

Name of Supervisor I :

Date : 30 JUNE 2025

Signature : _____

Name of Supervisor II :

Date :

Signature : _____

Name of Supervisor III :

Date :

SALES FORECASTING MODELS FOR DIRECT SELLING BUSINESS: A DATA-DRIVEN APPROACH TO PREDICTIVE ANALYTICS

SIVARAJAN A/L S.ESVARAN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master's in data science

School of Education
Faculty of Social Sciences and Humanities
Universiti Teknologi Malaysia

JUNE 2025

DECLARATION

I declare that this thesis entitled *"Sales Forecasting Models For Direct Selling Business: A Data-Driven Approach To Predictive Analytics"* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	.....
Name	:	SIVARAJAN A/L S.ESVARAN
Date	:	30 JUNE 2025

ACKNOWLEDGEMENT

In the course of preparing this thesis, I have had the privilege of interacting with many individuals, including researchers, academics, and industry practitioners, all of whom have contributed significantly to my learning and perspective. I would like to extend my heartfelt gratitude to my main thesis supervisor, Assoc. Prof. Dr. Mohd Shahizan Bin Othman, for his unwavering encouragement, insightful guidance, constructive criticism, and genuine support throughout this journey.

I am deeply grateful to my wife, as well as my father and mother, who have been my greatest source of motivation and strength, inspiring me to complete this master's degree and strive to make them proud.

I also wish to acknowledge my fellow postgraduate students for their support which made this journey more manageable and meaningful. My sincere appreciation goes out to all my UTM staffs and others who have provided assistance in various ways. Their valuable insights and suggestions have greatly enriched this work.

ABSTRACT

The purpose of this study is to develop and evaluate sales forecasting models for the direct selling business using transactional data from an Amway distributor in Malaysia. This study aims to assist independent distributors in improving inventory planning, target setting, and strategic decision-making by implementing data-driven forecasting approaches. The dataset comprised 553,542 sales transactions spanning from April 2023 to April 2025, covering detailed customer demographics, product sales, and transaction-level information. Data preprocessing included cleaning, outlier detection, and extensive feature engineering to create meaningful predictors such as temporal features, customer loyalty metrics, pricing indicators, and purchase recency measures. Four forecasting models were implemented and compared: Long Short-Term Memory (LSTM) neural networks, Random Forest, Linear Regression, and ARIMA. The results revealed that although LSTM, Random Forest, and Linear Regression models achieved high R^2 scores of 0.964, their high MAPE of 52.68% limited their practical utility for business forecasting. In contrast, ARIMA showed poor overall performance with a negative R^2 but paradoxically achieved 100% custom accuracy within the defined criteria. Overall, the findings highlight the need for careful model selection, robust evaluation frameworks, and further optimization to achieve reliable sales forecasting in direct selling businesses. This research provides valuable insights for distributors and contributes to advancing predictive analytics applications in the direct selling sector.

ABSTRAK

Tujuan kajian ini adalah untuk membangunkan dan menilai model ramalan jualan bagi perniagaan jualan langsung menggunakan data transaksional daripada pengedar Amway di Malaysia. Kajian ini bertujuan membantu pengedar bebas dalam meningkatkan perancangan inventori, penetapan sasaran, dan membuat keputusan strategik melalui pendekatan ramalan berasaskan data. Dataset yang digunakan mengandungi 553,542 rekod transaksi jualan dari April 2023 hingga April 2025, merangkumi maklumat terperinci tentang demografi pelanggan, jualan produk, dan data tahap transaksi. Pra-pemprosesan data melibatkan proses pembersihan, pengesanan outlier, dan kejuruteraan ciri secara meluas bagi menghasilkan pembolehubah peramal yang bermakna seperti ciri temporal, metrik kesetiaan pelanggan, indikator harga, dan ukuran kerecaman pembelian. Empat model ramalan telah dibangunkan dan dibandingkan: rangkaian neural Long Short-Term Memory (LSTM), Random Forest, Linear Regression, dan ARIMA. Hasil kajian menunjukkan bahawa walaupun model LSTM, Random Forest, dan Linear Regression mencapai skor R^2 yang tinggi iaitu 0.964, nilai MAPE yang tinggi pada 52.68% mengehadkan kegunaan praktikal model tersebut dalam ramalan perniagaan. Sebaliknya, model ARIMA menunjukkan prestasi keseluruhan yang lemah dengan nilai R^2 negatif tetapi secara paradoks mencapai ketepatan khas 100% dalam kriteria yang ditetapkan. Secara keseluruhannya, penemuan ini menekankan keperluan pemilihan model yang teliti, rangka kerja penilaian yang kukuh, dan pengoptimuman lanjut bagi mencapai ramalan jualan yang boleh dipercayai dalam perniagaan jualan langsung. Kajian ini memberikan pandangan berguna kepada para pengedar dan menyumbang kepada pembangunan aplikasi analitik ramalan dalam sektor jualan langsung.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARIMA	-	Autoregressive Integrated Moving Average
AutoML	-	Automated Machine Learning
BERT	-	Bidirectional Encoder Representations from Transformers
BGDM	-	Bass-Gumbel Diffusion Model
BiLSTM	-	Bidirectional Long Short-Term Memory
BLDM	-	Bass-Logit Diffusion Model
CNN	-	Convolutional Neural Network
CSV	-	Comma-Separated Values
DDM	-	Data-Driven Decision-Making
DNN	-	Deep Neural Networks
EDA	-	Exploratory Data Analysis
ELM	-	Extreme Learning Machine
ERP	-	Enterprise Resource Planning
EVs	-	Electric Vehicles
GRU	-	Gated Recurrent Unit
IQR	-	Interquartile Range
KNN	-	K-Nearest Neighbours
LSTM	-	Long Short-Term Memory
MAE	-	Mean Absolute Error
MAPE	-	Mean Absolute Percentage Error
MEMD	-	Multi-Angle Feature Extraction
MLP	-	Multi-Layer Perceptron
NEV	-	New Energy Vehicle
PDF	-	Portable Document Format
RFM	-	Recency, Frequency, Monetary
RMSE	-	Root Mean Square Error
RNN	-	Recurrent Neural Network
SVM	-	Support Vector Machine
TPOT	-	Tree-based Pipeline Optimization Tool

LIST OF SYMBOLS

R^2	-	Coefficient of Determination / R-squared
RM	-	Malaysian Ringgit
%	-	Percentage
|APE|	-	Absolute Percentage Error

LIST OF APPENDICES

APPENDIX	TITLE	PAGE

CHAPTER 1

INTRODUCTION

1.1 Overview

Direct selling, which is selling face-to-face outside of fixed retail locations, is a big business all over the world. Companies like Amway have grown their networks of independent distributors, who rely on personal connections and face-to-face deals to make sales. This model is flexible and lets you make changes and build relationships with customers, but it has some unique problems when it comes to inventory and business planning because consumer demand is very unpredictable and changes with the seasons (Al-maaitah, 2023; Korherr et al., 2022).

Predictive analytics is an important technology that helps people make strategic decisions in all areas of business in the data-driven economy. Companies use advanced forecasting models to figure out what will happen in the market, make sure they have the right amount of stock, and make customers happier. But the direct selling field, especially at the level of a single distributor, has not been quick to adopt predictive analytics and has instead relied on gut feelings instead of data-driven models.

The growth of machine learning and cloud-based analytics technology gives independent distributors a chance to completely change how they plan their sales. Distributors with good forecasting models can switch from a reactive business model to a proactive one, where they make the most of their business by managing their inventory strategically and based on accurate, predictable demand. This study looks at

how to create and use sales forecasting models for the direct selling industry, using the Amway distributor network as an example. The study aims to create useful, accurate forecasting tools for independent distributors to help them run their businesses better by using new predictive analytics techniques.

1.2 Problem Background

Independent distributors in direct selling companies have a particularly difficult time anticipating future sales performance and achieving successful inventory management. While it is common for larger retailers to have sophisticated enterprise resource planning (ERP) systems and an analytical team in place individual distributors rarely have access to these tools and forecasting capabilities (Daradkeh et al., 2022).

The nature of direct sales makes forecasting a nightmare. Sales trends can impact sales from various aspects including seasonal fluctuations, sales promotions, customer segment lifecycle changes, product launch, and economic environment. The first is that without predictive models to forecast these ebbs and flows, distributors either carry excess inventory that stifles cash or they go out of stock, stealing sales opportunities and angering customers.

Another issue is that the private and personal nature of person-to-person sales makes individual customer relations and purchasing behaviour within these diffuse sales force networks vary widely, thus preventing application of generic forecasting methods. The traditional time-series forecasting techniques may not be able to capture the complex dynamics involved in the direct selling relationships and need to be replaced with more complex models that can combine multiple factors and non-linear dynamics (Sivanathan et al., 2024).

Inaccurate sale forecast capability also affects strategic decisions at various hierarchies. Distributors face challenges in defining achievable monthly and quarterly targets, or preparing parties, and deciding how to further develop or focus their market segments. And their inability to predict and respond proactively often means they're unable to optimize their monetization opportunity and build a solid, growing business.

1.2 Problem Statement

While it is a well-known fact that predictive analytics have brought tremendous success in the retail and e-commerce industries, individual distributors in direct selling companies do not have access to advanced sales forecasting models that can forecast future performance, steering strategic business decisions. Without accurate forecasting facilities suboptimal business results inefficient stock management, missed sales opportunities, poor planning for promotions, the list goes on. Thus, the fundamental problem solved by this work is as follows:

How to design and implement advanced sales forecasting models to generate reliable, actionable predictions to direct selling industry, allowing independent distributors to streamline their operations with data-based prediction analytics?

1.4 Research Questions

Based on the comprehensive analysis of Amway transaction data spanning April 2023 to April 2025, this research addresses the following key questions:

1. What are the dominant temporal patterns in direct selling transactions, and how do seasonal variations affect sales forecasting accuracy across different time horizons?
2. How do traditional statistical models (ARIMA) compare to machine learning approaches (LSTM, Random Forest, Linear Regression) in terms of forecasting performance for direct selling businesses with highly variable sales patterns?

3. What factors contribute to the superior performance of ARIMA models in achieving acceptable forecasting criteria compared to machine learning approaches in this specific business context?
4. How can the identified customer demographics patterns and purchasing behaviours be leveraged to improve sales forecasting models?

1.5 Research Aim and Objectives

Aim: To investigate and evaluate the effectiveness of different sales forecasting methodologies in direct selling business environments through comprehensive data analysis and model comparison, in order to determine the most suitable approaches for independent distributors facing highly variable market conditions.

Objectives:

- **To conduct comprehensive exploratory data analysis** of transactions over a 24-month period to identify key patterns, trends, and characteristics that influence sales forecasting in direct selling environments.
- **To analyse temporal sales patterns** including monthly seasonality, day-of-week effects, and yearly trends to understand the cyclical nature of direct selling transactions and their impact on forecasting model selection.
- **To develop and implement multiple forecasting models** including traditional time series methods (ARIMA) and modern machine learning approaches (LSTM, Random Forest, Linear Regression) to address different aspects of sales prediction challenges.
- **To establish comprehensive model evaluation criteria** using multiple performance metrics to provide robust assessment of forecasting effectiveness across different business contexts.
- **To identify optimal forecasting approaches** by comparing model performance against established business criteria to determine practical applicability for direct selling operations.

1.6 Research Scope

This study will use detailed transaction and customer data from a single Amway distributor over the period April 2023 to April 2025. The study will address the development of forecast models from short-term (weekly) to long-term (quarterly) forecast horizon.

Critical components of the project scope are as follows:

- Data Sources: Two year of granular sales transaction data, customer demographics, product catalogues, promotional calendars, and external factors such as economic indicators, seasonality.
- Implementing and comparing traditional statistical forecasting models like ARIMA, Exponential Smoothing and more modern machine learning models like Random Forest, **Linear Regression**, LSTM).
- Horizons of Prediction: Developing systems that can model predictions at representative time scales, 1-week, 4-weeks, and 12-weeks, to meet different business planning requirements.
- Tech Stack: Python for model-building, focusing on scikit-learn, Tensorflow/Keras, as well as specialized forecasting libraries like Prophet and statsmodels.
- Validation Framework: Develop robust cross validation methods and walk-forward analysis to ensure that the model was reliable and did not suffer from overfitting.
- Deployment Consideration: Develop models considering deployment challenges, such as automation, scale and ease of use for non-technical users.

1.7 Significance of Research

This research contributes significant value to both academic knowledge and practical business applications. On a business level, it fills a unique void within the direct selling industry as it allows independent distributors access to a level of predictive analytics they would not otherwise have access to. The power of being able to accurately predict sales can revolutionize the way that a distributor operates, driving more efficient inventory management, better customer service and greater business profitability. From the perspective of the wider direct selling industry, a system for scaled deployment of predictive analytics is proposed through this research. Companies such as Amway can use these findings to design tools to support distributors, better training programs, and to improve network effectiveness in general. The methods we have built are a potential template for other direct selling businesses to update their analysis capabilities.

Theoretically, this paper adds to the burgeoning field of retail analytics by investigating distinct challenges and opportunities in direct selling settings. It contributes to knowledge regarding the uses of traditional forecasting techniques for person-to-person sales and offers practical applications of machine learning for small business. The study also adds to the more general discussion of democratization of advanced analytics, revealing that complex predictive modelling tools can be made available for single entrepreneur and small business owner level players. This has implications beyond direct sales and potentially contribute to the design analytics tool for other types of small business sectors. In addition, the research helps advance knowledge in sales forecasting and predictive modelling, related to the efficiency of various forecast methods in the largely unstable, relationship-driven business context.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents a complete review of literature, which provides the theoretical basis for the development of a series of advanced sales forecasts models dedicated for direct selling industry. The review covers the literature in the various fields for sales forecasting techniques, application of predictive analytics in retail and direct selling, machine learning techniques in business forecasting, and the multi-source data analytic in entrepreneurial engagement.

This literature review is organized to consider the distinctive challenges that face direct selling organizations, particularly those functioning within a network marketing type of relationship such as Amway and other multi-level marketing groupings. DSS in direct selling businesses are less technologically advanced when compared with traditional retail systems as they cover extensive distributed network of independent distributors and the decision-making based on advance forecasting techniques and data driven approach is not commonly available.

This paper aims to provide an overview and assessment of predictive analytics techniques that can be deployed in the direct selling domain and applied to the prediction of sales pattern one that takes into consideration the distributor performance. Through the integration of research from the related fields such as retail forecasting, customer relationship management, business intelligence, this review attempts to provide a theoretical guideline for constructing sound sales forecasting model.

This chapter reviews and systematically analyses previous works in the literature to see how advanced predictive analytics, such as machine learning algorithms, time series forecasting strategies, and hybrid ensemble computing, can be applied to deal with these challenges. In particular, the review highlights data-driven approaches tailored to cope with the specific features of direct selling models such as the nature of sales organization, irregular sales patterns, distributor churn, network growth and multi-level effect.

2.2 Overview of the Direct Selling Industry

Direct selling is a retail channel used by top global brands and smaller, entrepreneurial companies to market products and services to consumers. Independent distributors or agents serve as intermediate links between the company and the final customer. One of the most well-known of these models is Amway, which has approximately 3 million independent distributors worldwide and operates in more than 100 countries (Mondom, 2018). Contrary to usual employment, Amway believes in a concept of "independent business ownership" allowing people to become entrepreneurs with adjustable size of investment and low entry costs. Distributers are supported to plan their own time and business and give their personalised service mostly in a similar network of people and friends, rather than random or door to door selling (Mondom, 2019).

Amway works on a multi-level marketing plan, and people who distribute its T-shirts and eye cream not only sell the products, but also are encouraged to build a network, on which new levels of sellers bring bonuses to the upline based on the downline's sales. Compensation is commission-based and, for some products, the more you sell and sponsor the higher you work up the ranks through a set of metal-themed ranks that describe the size of your sales base and the size of your network (Mondom, 2018).

More than just a seller of goods an array that spans from household cleaners to health supplements and cosmetics, Amway peddles a larger philosophy, “compassionate capitalism,” with its emphasis on self-empowerment, self-reliance and financial independence. Amway’s founders, Richard DeVos and Jay Van Andel, presented the company as part of a so-called life liberation movement for people who shun traditional jobs and yearn for more control over their financial and personal destinies. This philosophy mixes free enterprise with cooperation and mutual aid while encouraging distributors to be independent businesspeople, not government or company welfare cases (Mondom, 2018; Mondom, 2019).

Amway's model is appealing, but it's not an easy road for distributors. They frequently have few formal business resources, training or sophisticated analytical tools at their disposal, relying on people skills such as charm and persuasion to develop and maintain relationships with customers, and attract new salespeople. Most distributors work with those that do not necessarily equate with centralized data systems or sales tracking that is full, this often causes dis-economies in terms of inventory and promotion planning. The pressure to replicate their upline's sales and recruiting process can however cause pressure and frustration among beginning reps, which can consequently foreclose or discourage newcomers wishing to join the business. Yet the hybrid corporate structure of Amway, taking facets of small independent entrepreneurship and the resources and discipline of a large corporation, has allowed the company to flourish as it opened operation in dozens of countries, turning its founders into major figures not just in business but also in political conservatism and the promotion of free enterprise (Mondom, 2018; Mondom, 2019).

2.2.1 Challenges in the Direct Selling Business Model

The direct selling model has some of its own operational challenges at the individual distributor level, despite its flexibility and scalability. There are few formal commercial resources available to distributors, such as advanced training and tiny analytics, which may lead to suboptimal sales and business optimization strategies.

Rather, the rareness of these commodities leaves most of the distributors to depend primarily on their personal abilities appeal, influence and relationship building to advance sales, enlist prospects and build a business. A major problem is the fact that there is no database systems, no seller has access to a central database or complete transaction data to help analyse and forecast sale. This lack leads to a suboptimal inventory management; the stocking policy is more a matter of guess than a precise estimation of demand and it may lead to stockouts or overstocks (DeCarlo et al., 2025).

Another huge issue is bad promotional timing, where marketing efforts are usually 'reactive' rather than 'proactive', failing to take into account past statistics to optimize campaigns. These weaknesses are complemented by intra-organizational processes - or as the studies say it, "the last mile" internal selling - where organizational salespeople are confronted with numerous levels of hierarchical approval and by highly specialized forces. Deals take forever to get approved as salespeople have to play Tetris with layers of management and the sheer inertia can cost sales in fast-moving markets. Interacting with dedicated individuals in other departments and interrupt-driven communication processes can result in fragmented communications and inconsistent processes which hinder the close of sale. Such bureaucratic frustrations can leave salespeople feeling disenchanted and they disengage, working on its bigger customers who are more likely to get their order rushed through at the expense of smaller but possibly more profitable clients. Added together, these challenges often result in high new distributor attrition and reduced long-term profitability and growth in the direct selling business model (DeCarlo et al., 2025).

2.3 Role of Data Analytics in Business Decision-Making

Traditionally the formulation of business decisions predominantly relied upon experiential knowledge, intuitive judgment, and anecdotal evidence. Nevertheless, the advent of digital transformation has progressively directed organizations towards the paradigm of data-driven decision-making (DDM), which underscores the necessity of making informed selections predicated on empirical data, statistical methodologies,

and predictive analytical frameworks. This transformation culminates in augmented operational efficiency, heightened customer satisfaction, and superior financial performance (Colombari et al., 2023; Szukits, 2022). Within the realm of small-scale entrepreneurship, inclusive of direct selling, DDM empowers individual sellers to discern lucrative product categories, comprehend consumer purchasing patterns, project demand with greater precision, and refine marketing strategies. However, the proliferation of DDM remains inadequate among independent distributors, attributable to deficiencies in technical competencies, restricted access to requisite tools, and organizational impediments concerning the effective interpretation and utilization of data (Colombari et al., 2023; Szukits, 2022).

Scholarly investigations emphasize that efficacious DDM transcends mere data availability; it fundamentally relies on an organization's capacity to seamlessly integrate and meaningfully process heterogeneous data streams (Colombari et al., 2023). Organizations must cultivate a digital orientation—characterized as a firm's dedication to harnessing digital technologies—to optimize the utilization of advanced analytical techniques, which subsequently enhances decision-making processes (Szukits, 2022). The function of integrators, such as controllers who amalgamate analytical proficiency with business acumen, is indispensable in transforming intricate data into actionable insights that managerial personnel can comprehend and implement (Szukits, 2022). Nonetheless, even in instances where information is accessible, decision-makers do not invariably depend solely on it; intuition and experiential knowledge continue to exert substantial influence, particularly when confronted with ambiguous or complex decision scenarios (Colombari et al., 2023; Szukits, 2022). Consequently, the transition from intuition-based decision-making to data-driven methodologies is multifaceted and contingent upon organizational preparedness, technological infrastructure, and human factors that facilitate the proficient interpretation and application of data.

2.4 Customer Segmentation and Behaviour Analysis

Customer segmentation is the practice of dividing a massive market into smaller subgroups according to demographics, purchasing behaviour and product preferences, among others. This strategy is essential for businesses aiming to achieve more, personalized and targeting marketing messages, increased retention through personalized engagement, efficient use of resources and, in the end, enhancing customer satisfaction. For independent distributors in direct selling, it means better segmentation of high value customer segments, more efficient targeting, lower spam rates, increased conversion rate and, in general, increased sales. Recent advances in artificial intelligence and machine learning have revolutionized customer profiling and segmentation. Methods include Recency, Frequency, and Monetary analysis combined with clustering algorithms, such as K-means. As an example, best customers, new customers, and intermittent customers may be identified. They should be treated differently to maximize engagement and loyalty (Kasem, Hamada, & Taj-Eddin, 2024). It fosters the use of data-driven predictions and placement of all effort into likely high potential groups. Secondly, models improve on accuracy as they don't discard low-priority correlations automatically as with man-based intuition. It improves sales efficiency and the strength of customer ties through personalized and relevant engagements.

2.4.1 Behavioural Insights and Personalization

It is important for sellers to understand customers and their buying behaviours to meet their needs. Important metrics such as how often a customer places an order, the average order value, the return rate, and the intensity of communication through platforms such as WhatsApp or social media serve as critical data for analysing consumer trends and behaviours in terms of their preferences and satisfaction desires. (Zhou et al., 2025). By using big data algorithms, retailers can build recommendation systems that facilitate a personalized shopping experience for customers which increases their retention ratio and marketing conversion. For example, Amazon and

Netflix gain knowledge from the customers' buying and viewing decisions and deduce predictions to present to the customers their future preferred content (Zhou et al., 2025). This subsequently increases their overall revenue and retention ratio since the customer gets what they prefer. Focusing on behaviour, they can make offers and use personalized pricing where they gauge the effects of price elasticity on the customers, thus, optimizing their revenue and the consumer comfort data. As much as it gives a competitive advantage, it brings privacy concerns which are likely to affect the collective. Sellers need to bear in mind the customer's preference and develop a policy that ensures the customer opts in the information sharing principle (Zhou et al., 2025). In conclusion, integrating behavioural science in earning strategies is a critical stage in leaving the went fortuity behind in earning decision but replaces it with specific, processed offers and retention choices.

2.5 Product Performance and Return Pattern Analysis

2.5.1 Measuring Product Performance

Evaluating product performance is paramount for discerning which items significantly enhance a corporation's revenue, profit margins, and overall consumer satisfaction. The prevalent metrics employed encompass sales volume per product, revenue contribution, profitability index, customer feedback scores, and rates of return or refund. Recognizing top-performing products allows sellers to effectively prioritize inventory management and marketing initiatives, while products that exhibit subpar performance may necessitate revaluation or discontinuation to optimize resource allocation and profitability. Investigations into adaptive selling and personal selling underscore that the performance of sales personnel has a substantial impact on product success; seasoned sales professionals are adept at customizing their strategies to align with customer needs, thereby augmenting product sales and consumer satisfaction.

Adaptive selling, characterized by the alteration of sales behaviour to accommodate varying customer circumstances, exerts a positive influence on salesperson performance and, by extension, product performance through enabling sales personnel to respond adeptly to a wide array of customer demands. Furthermore, personal selling methodologies that emphasize persuasive communication and customer education play a critical role in enhancing product acceptance and fostering loyalty. The experience accrued through sales practice enhances the capability of sales personnel to differentiate products and deliver value to consumers, thereby reinforcing the correlation between effective sales strategies and improved product performance metrics (Rianita, 2022).

2.5.2 Analysing Return and Refund Patterns

Returns are a huge issue for sales companies, including direct sellers where buyers often cannot touch or try on a product before they purchase it. A high level of returns is an indicator of several issues including discrepancies in product indulgence between the advertising and the customer expectation, poor product quality, misaligned pricing strategy, and lack of customer knowledge (El Kihal, Erdem, Schulze, and Zhang, 2025). Sellers can also identify what is driving returns and develop targeted ways to increase customer satisfaction and reduce return rates by analysing return patterns by product category, customer demographics and shopping time. Studies have demonstrated that customer return rates increase over time and post-purchase return behaviours become habitual which can potentially override brand engagement and product familiarity benefits (El Kihal et al., 2025). The research points out, that if customers have a record of high return rates, they have a high probability of returning products, and this would result in the formation of a ‘return habit’, which retailers should bear in mind. Brand experience can decrease returns by making it easier for consumers to feel comfortable that a product is going to work for them and be of high quality. But this tendency is usually dwarfed by habits of long-term return, which begets more returns when ever we buy.

Plus, pricing dimensions matter a lot when it comes to returns. The more an item costs, the higher its rate of return, and the cheaper the item, the lower its return rate. Age and sex are also key demographics that impact on return patterns. For instance, older customers have lower return rates, while female customers exhibit higher return rates (El Kihal et al., 2025). These findings illustrate the critical need for direct sellers or retailers to monitor and analyse return data constantly. This will enable them to enhance their product selections, customer communications strategies and rules that shade heavily toward forgiveness to profitability, which in turn will enable them to establish longer-term relationships with their customers.

2.6 Sales Forecasting Models

Sales forecasting has evolved significantly with advances in computational methods and data availability. Traditional approaches can be categorized into statistical models, machine learning models, and hybrid decomposition-ensemble frameworks.

2.6.1 Traditional Statistical Models

Classical forecasting approaches include time series models such as ARIMA, exponential smoothing, and regression-based methods. Bass diffusion models remain prevalent for new product sales forecasting, with recent extensions including the Bass-Gumbel diffusion model (BGDM) and Bass-Logit diffusion model (BLDM) demonstrating improved performance for products with seasonal effects (Cosguner & Seetharaman, 2022; Fernandez-Durán, 2014).

Grey models have shown particular effectiveness for small datasets and low-frequency data. Recent developments include self-adaptive optimized grey models and time-varying grey Bernoulli models, which have been successfully applied to electric vehicle sales prediction (Ding & Li, 2021; Zhou et al., 2023). These models address the challenge of limited historical data while maintaining computational efficiency.

2.6.2 Machine Learning Approaches

The adoption of machine learning for sales forecasting has gained traction thanks to better performance in high-dimensional data and non-linear relationships. SVM and ELM have been proved to be resistant in different retail environments (Chen & Zhao, 2024; Zhang et al., 2023). Deep learning methods, such as the Long Short-Term Memory (LSTM) networks and their bidirectional versions (BiLSTM), have gained a lot of attention in learning temporal dependencies in sales data. Recent models present hybrid CNN-LSTM models for neural energy vehicle sales prediction which effectively aggregates spatial feature extraction with temporal sequence modeling (Li et al., 2024a; Wang, 2022).

Automated Machine Learning (AutoML) is a big breakthrough in democratizing Predictive Analytics. TPOT (Tree-based Pipeline Optimization Tool) and other similar toolkits automate the search for the best combination of features, model configurations (feature generation) and parameter settings, democratizing advanced forecasting to those who are not experts in machine learning (Olson & Moore, 2016; Alsharef et al., 2022).

2.6.3 Various Machine Learning Models

Long Short-Term Memory (LSTM)

LSTM networks are a kind of RNN architecture explicitly devised to handle the sequential nature of the data and to capture long-term dependencies overcoming the vanishing gradient problem that is still a common issue in classic RNNs. In sales prediction, we have applied LSTM models extensively to identify seasonality, and trend shifts throughout historical sales data. Yan et al. (2025) also proposed a new sales prediction framework by employing LSTM as an estimator to capture time-related features within a separated-by-feature-extraction module and demonstrated that isolating sequential features from static-features prevents the downgrade of model's

accuracy caused by the blend of features. Similarly, Liu et al. (2025) has employed LSTM models on hybrid models to forecast electric vehicle sales by incorporating BERT- BiLSTM-based sentiment analysis with decomposition techniques for better representing complex multiscale and nonlinear sales data. While LSTM models are effective at learning temporal dynamics, they are data hungry and may be overfit at will even without proper regularisation techniques.

Random Forest

Random Forest is an ensemble learning model which creates a set of decision trees based on randomly subsampled training data and averages their prediction to increase accuracy and reduces overfitting. In sales prediction studies, Random Forest achieved good performance of dealing with non-linear relationships and high-order conjunctions of multiple features. For instance, Rahman et al. (2025) used Random Forest for predicting sales in supply chain and found that though the performance of the Random Forest model was good, Voting Regressor (combination of Random Forest and other models) best accuracy with RMSE of 1.54 and R^2 0.9999, score over the base models. Random Forest modelling is resistant to outliers and multicollinearity and provides insights about the importance of the features that can be useful for business. They do not, however, naturally capture the order of the temporal dependencies and are not straightforward to use as predictive models for time series data without the addition of lagged features or decomposition-based preprocessing.

ARIMA

It is a traditional statistical technique for univariate time series forecasting. It is a combination of autoregression (AR) and moving average (MA), where autoregression is calculated on the differenced data and the moving average is calculated on the errors. In the literature reviewed, ARIMA and its extension, ARIMAX (ARIMA with exogenous variables), are heavily applied for short range sales forecasting due to their interpretability and good fit for the stationary linear

patterns. Elalem et al. (2023) analysed ARIMAX and deep neural networks in new products with short life cycles sales forecasting, showing that in clean data scenarios ARIMAX was advantageous over DNNs, while in noisy data scenarios, DNNs were more robust. Even though ARIMA has advantage in modelling the time series, it is unable to capture the nonlinear patterns or multivariate relationships as in direct selling and modern retail data.

Linear Regression

Linear Regression is one of the basic statistical regression models that models the relationship between a dependent variable and one or more independent variables using a linear equation. In sales forecasting research, the Linear Regression is usually a benchmark model in that it is simple, interpretable, and computationally efficient. Rahman et al. (2025) also adopted the Linear Regression for supply chain demand prediction in their comparative study where they pointed out that despite of the quick preliminary insights given by the model, 3 it was outperformed by the more sophisticated machine learning models, as Random Forest and Voting Regressor, which take into account complex information structures. The linear relationships and independent observations assumed by the model restrict its predictive accuracy when the data exhibit nonlinearities, interactions or temporal dependencies that are not pre-processed.

Table 2.1 Previous Work on Sales Forecasting

Author / Year	Title	Research Focus	Machine Learning Methods
Liu et al. (2023)	A combination model based on multi-angle feature extraction and sentiment analysis: Application to	Developing a hybrid forecasting model integrating multi-angle feature extraction and sentiment analysis	MEMD decomposition, Sentiment analysis, Combination forecasting

	EVs sales forecasting	for electric vehicle sales prediction	
Elalem et al. (2023)	A machine learning-based framework for forecasting sales of new products with short life cycles using deep neural networks	Forecasting sales of new short life cycle products using deep learning and ARIMAX with cluster-based data augmentation	ARIMAX, LSTM, GRU, CNN
Yan et al. (2025)	A novel sales forecast framework based on separate feature extraction and reconciliation under hierarchical constraint	Hierarchical sales forecasting with separate feature extraction and reconciliation to improve supply chain planning	LSTM (for time-dependent features), MLP (for static features)
Liu et al. (2025)	An electric vehicle sales hybrid forecasting method based on improved sentiment analysis model and secondary decomposition	Combining sentiment analysis and secondary decomposition for electric vehicle sales forecasting	BERT-BiLSTM sentiment analysis, decomposition + ML hybrid

Wu et al. (2023)	Bayesian non-parametric method for decision support: Forecasting online product sales	Developing PoissonGP, a Bayesian non-parametric model for online sales forecasting with uncertainty quantification	Poisson Gaussian Process (PoissonGP)
Rahman et al. (2025)	Enhancing sustainable supply chain forecasting using machine learning for sales prediction	Using ML algorithms to improve demand prediction and supply chain decision-making	Linear Regression, Elastic Net, KNN, Random Forest, Voting Regressor
Hu et al. (2025)	Grid-based market sales forecasting for retail businesses using automated machine learning and geospatial intelligence	Combining AutoML and geospatial intelligence for grid-level market sales forecasting and site selection	AutoML, regression models
Shao et al. (2025)	New energy vehicles sales forecasting using machine learning: The role of media sentiment	Integrating media sentiment indices into machine learning models for NEV sales forecasting	ML models with sentiment analysis (exact algorithms not detailed but includes ML hybrid models)

2.7 Research Gap

Recent sales forecasting literature has identified a number of key limitations that restrict the wider adoption and effectiveness of existing approaches. First, a key limitation is spatial selectivity, for most studies are specific to markets in one or two countries, notably China and the United States. This emphasis limits the applicability of the results to other areas with distinct cultural patterns, economic situations, and regulatory frameworks, while they may not be readily transferable onto the field of emerging markets or consumer behaviours.

Another limitation of the literature is that there has been little attempt to explore time-wise dynamics like the magnitude of importance of different predictor factors such as sentiment indicators, aspatial data and economic indicators varies across different forecasting horizons, from very short-term operational planning to the long-term strategic decision.

Furthermore, comparatively lack of research about practical challenges of real-time data integration, particularly the integration of streaming sentiment data and market signals into forecasting systems without compromising computational efficiency and forecasting accuracy.

Finally, there has been little research on multi-modal analysis aimed at unifying various types of data, such as text from social media or news, visual contents from marketing or user-generated content, and the traditional numerical data, in a coherent manner into the same forecasting workflow. This is a huge opportunity to enhancing prediction capabilities through comprehensive data integration.

2.8 Summary

This chapter includes a comprehensive literature review of ongoing research regarding sales forecasting models and predictive analytics specifically tailored for direct selling businesses. This chapter presents an analysis of the similarities and differences between various forecasting methods, machine learning algorithms, and data-driven approaches used in retail and network marketing contexts. Apart from that, this chapter also provides an in-depth discussion regarding the direct selling industry characteristics, particularly focusing on the Amway business model and the unique challenges faced by independent distributors. The next chapter will discuss the research methodology and outline the main strategies used in developing advanced sales forecasting frameworks for direct selling businesses.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The research methodology employed is discussed in this chapter to develop an accurate sales forecasting model for the direct selling businesses. The methodology outlines the process of collecting data from the primary direct selling website, followed by data preprocessing, feature engineering, model development, and evaluation using machine learning techniques to forecast future sales performance. This study aims to generate meaningful insights from historical sales data and external factors that influence sales in direct selling environments. Ultimately it will provide valuable forecasting capabilities for business planning and decision making.

3.2 Research Framework

The framework for this research includes the following stages:

1. Identifying the Research Problem and Reviewing Existing Literature.
2. Data Collection from Amway Business Operations.
3. Preprocessing the Data: Preparing and cleaning data for detailed analytical tasks
4. Exploratory Data Analysis (EDA): Time Series Analysis and Pattern Recognition
5. Sales Forecasting Models: Implementing multiple forecasting algorithms (Linear Regression, Random Forest, LSTM and ARIMA)

6. Model Evaluation: Comparing model performance using forecasting evaluation metrics.

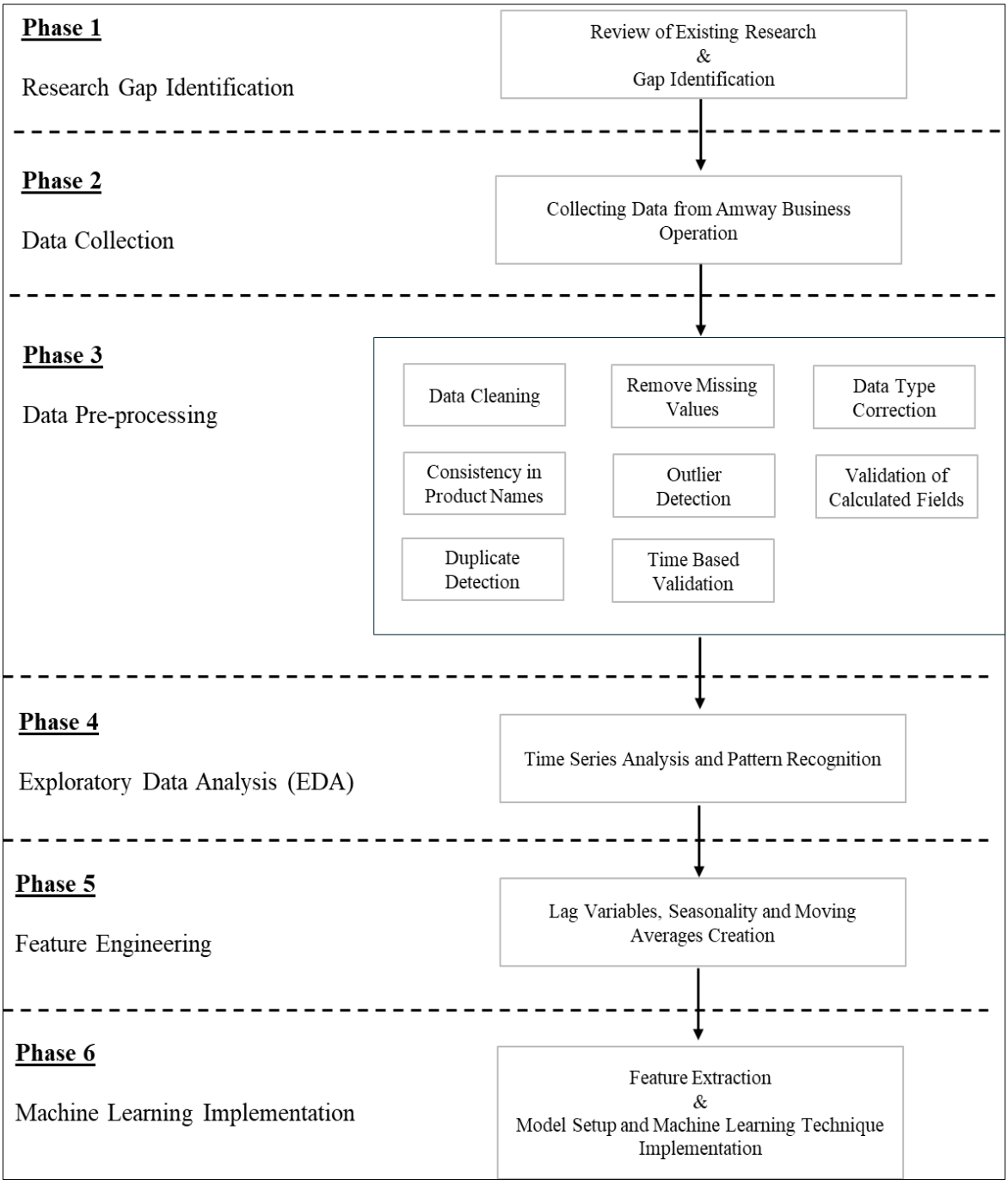

Figure 3.1 Research Framework for a Sales Forecasting Model in Direct Selling Business

3.3 Problem Formulation

This study aims to build a precise sales forecasting system for direct selling distributors using machine learning techniques. Thus providing valuable insights for strategic planning. However. Ensuring precise and dependable forecasting requires tackling a number of key issues:

a) Identifying key factors that influence sales performance in direct selling business.
b) Handling seasonality, trends and cyclical patterns in sales data.
c) Comparing the performance of Random Forest Linear Regression, LSTM and ARIMA algorithms in sales forecasting.
d) Developing robust model that can adapt to changing market conditions.

3.4 Data Collection

Data was collected from an individual Amway distributor sales records to develop a sales forecasting system that helps distributors to enhance their business growth and sales performance. The dataset represents actual transaction level sales data from an Amway distributor's business operations. This sales data provides detailed insights into customer purchasing patterns, product performance and sales trends.

Dataset Information:

1. **Collection Period**: April 2023 to April 2025. (24 consecutive months)
2. **Data Format**: Monthly PDF sales reports downloaded from the distributor portal.
3. **Conversion Process**: Python-based PDF to CSV conversion.
4. **Variables:** 12 comprehensive features per sales record
5. **Final Dataset**: 553,542 sales records in structured CSV format.

3.4.1 PDF to CSV Conversion Process

Python-based Conversion Methodology

The conversion from PDF to CSV format was accomplished using Python libraries specifically designed for PDF data extraction and processing.

Step 1: Environment Setup

```
python# Required libraries for PDF to CSV conversion
import pandas as pd
import PyPDF2
import tabula
import pdfplumber
import os
import glob
from datetime import datetime
import re

# Install required packages
# pip install pandas PyPDF2 tabula-py pdfplumber openpyxl
```

Figure 3.2 Importing Modules for PDF Handling

Step 2: PDF Data Extraction Function

```
python
def extract_amway_sales_from_pdf(pdf_file_path):
    """
    Extract sales data from Amway monthly PDF reports
    """
    try:
        # Method 1: Using tabula-py for table extraction
        tables = tabula.read_pdf(pdf_file_path,
                                 pages='all',
                                 multiple_tables=True,
                                 pandas_options={'header': 0})

        # Combine all tables from the PDF
        if tables:
            combined_df = pd.concat(tables, ignore_index=True)
            return combined df
```

Figure 3.3 Function to Extract tables from Sales PDF Reports

Step 3: Batch Processing All Monthly PDFs files

```
python
def process_all_monthly_pdfs(pdf_folder_path):
    """
    Process all monthly PDF files and combine into single dataset
    """
    all_monthly_data = []
    pdf_files = glob.glob(os.path.join(pdf_folder_path, "amway_sales_*.pdf"))

    print(f"Found {len(pdf_files)} PDF files to process")

    for pdf_file in sorted(pdf_files):
        print(f"Processing: {os.path.basename(pdf_file)}")

        # Extract data from current PDF
        monthly_data = extract_amway_sales_from_pdf(pdf_file)
```

Figure 3.4 Function to Process and Combine Monthly Sales Reports

Step 4: Export to Final CSV

```python
def export_to_csv(data, output_filename='Amway_Sales_Records_100k.csv'):
    """
    Export processed data to CSV format
    """
    # Final data validation
    print("Final dataset summary:")
    print(f"Total records: {len(data)}")
    print(f"Columns: {list(data.columns)}")
    print(f"Date range: {data['Date'].min()} to {data['Date'].max()}")

    # Export to CSV
    data.to_csv(output_filename, index=False)
    print(f"✓ Data exported to {output_filename}")
```

Figure 3.5 Function to Processed Sales Data to CSV

Number of rows: 553542
Number of columns: 12

	Order_ID	Date	Time	Customer_ID	Product_ID	Product_Name	Quantity	Unit_Price	Total_Amount	Return_Status	Customer_Age	Customer_Name
0	1000536015	2023-04-01	00:01:00	7010445678	125895A	Hand Sanitizer 400ml	5	59.0	295.0	No	36	Murugaiah A/L Ahyanari
1	1000526177	2023-04-01	00:04:00	7010045678	121697	DOUBLE X Refill Pack 186tab	3	198.0	594.0	No	45	Manirajah A/L Velu
2	1000045590	2023-04-01	00:06:00	7009583801	126457	Anti-Hair Fall Shampoo 280ml	2	72.0	144.0	No	39	Santhi A/P Sinnkoladai
3	1000119832	2023-04-01	00:07:00	7013409072	230727	Sanita Ultra Thin Wings 20/pk	5	13.0	65.0	No	43	Sumathi A/P Supaya
4	1000246702	2023-04-01	00:07:00	7024498970	102735	ClearGuardâ„¢ 180tab	4	139.0	556.0	No	44	Sheela A/P Berabakaran
...	...	...	...	...	...	...	...	...	...	...	...	...
553537	1000926116	2025-04-30	23:46:00	7010389012	123785	Renewing Reactivation Cream 50ml	4	250.0	1000.0	No	54	Perabakaran A/L Ramasamy
553538	1000234084	2025-04-30	23:48:00	7013658426	309177	Vergold Drip Coffee 10 sachets x 11g (Medium R...	5	58.0	290.0	No	39	Yugneswari A/P Rajendran
553539	1000076280	2025-04-30	23:49:00	7010045678	387800	Pursueâ„¢ Disinfectant Cleaner One Step 1l	3	30.8	92.4	No	45	Manirajah A/L Velu
553540	1000829916	2025-04-30	23:56:00	7686255	319372M	White Tea Toothpaste 200g	4	29.0	116.0	No	51	Tamil Selvi A/P Velayutham
553541	1000915819	2025-04-30	23:57:00	7010156789	592300	Garlic with Licorice 150tab	4	97.0	388.0	No	54	Arjunan A/L Pachappan

Figure 3.6 The Dataset Preview

3.5 Data Pre-Processing

Conducting an initial analysis is essential before proceeding with further preprocessing to ensure a thorough understanding of the dataset's feature availability. Several data processing and transformation procedures will be applied to prepare the data for time series forecasting.

Data Pre-Processing	Purpose
Preliminary Analysis	To analyse the given dataset and extract meaningful insights that will support the subsequent modelling phase.
Data Cleaning	Remove Missing Values, inconsistent records and outliers.

Table 3.1 Data Pre- Processing Method

3.5.1 Preliminary Analysis

Preliminary analysis plays a vital role in any data analysis project, which helps for a good understanding of the dataset, including its variables, format and structure. This preliminary identification helps to uncover potential issues. For example, outliers, inconsistencies and missing values need to be identified to ensure reliable and accurate results.

The Preliminary Analysis on this case involves two key stages:

a) Identifying common patterns within the raw data.
b) Analyzing data distribution based on time and relevant keywords.

3.5.2 Data Cleaning

In sales forecasting, data cleaning plays a vital role in guaranteeing that the data is precise, pertinent, and ready for processing by predictive models. The following steps outline the data cleaning procedures applied on the Amway Distributor Sales Dataset.

1. **Remove the Missing Values**

   Missing values are important to identify any rows and columns where data is missing such as blank entries. These missing values can lead to incorrect analysis or errors during modelling. Therefore, it's essential to remove rows or

columns with critical missing information or fill non-critical ones using reasonable estimates.

2. **Data Type Correction**

   Convert data into correct formats for example, date strings to datetime or text numbers to integers/floats. A proper data types allow for accurate calculations and visualizations.

3. **Consistency in Product Names and ID's**

   This to ensure product names in IDs are standardized across all the sales records. Inconsistent naming such as "G&H Lotion" vs "GH Lotion" leads to incorrect grouping. To address this, replace typos manually or use string normalization techniques.

4. **Outlier Detection**

   This steps involve to finds extreme or unrealistic values such as negative sales and absurd quantities. Outliers can skew analysis and mislead insights. To manage it, statistical methods (IQR, Z-scores) can be applied or simple filtering such as ensuring quantities are greater than zero can also be used.

5. **Validation of Calculated Fields**

   This step ensures calculated fields like Total Sale Amount reflect expected values (Quantity x Unit Price). Miscalculations can lead to inaccuracies in revenue analysis. To prevent this, recalculate and compare with any mismatches flagged for review or corrections.

6. **Duplicate Detection**

   This step focuses on identifying and removing duplicate rows for example transactions have been recoded twice. Duplicates can distort sales counts and totals, which leads to inaccurate analysis. Use drop duplicates () based on key identifying columns like Product ID to ensure data accuracy.

7. **Time Based Validation**

This validation method is to check whether dates fall within valid ranges. For example, ensuring they are not in the future unless they represent pre-orders. The invalid dates can affect time series analysis. This will be done by each date against the current date or a defined date range.

8. **Apply Pre-Processing**

This step integrates all the previously mentioned procedures into a single preprocessing pipeline, ensuring that the entire dataset is cleaned and processed consistently and uniformly.

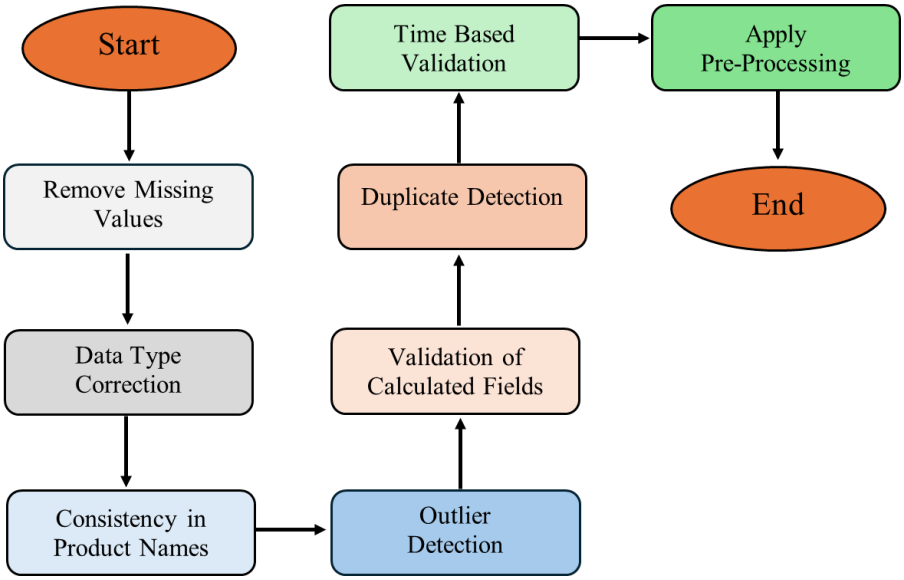

Figure 3.7 Flow Data Cleaning and Preparation

In this section, data cleaning is performed to identify and eliminate rows and columns with missing values. As illustrated in Figure 3.7, the data pre-processing steps include data type correction, standardizing the product names, outlier detection, validating calculation fields, normalizing, duplicate detection and time-based validation.

```python
import pandas as pd
import numpy as np

# Load Amway sales dataset
df = pd.read_csv('Amway Sales Records_100k.csv')

# Execute complete cleaning pipeline
print("Starting Amway data cleaning pipeline...")

# Apply all cleaning steps
df = remove_missing_values(df)
df = correct_data_types(df)
df = standardize_product_names(df)
df = detect_outliers(df)
df = validate_calculations(df)
df = remove_duplicates(df)
df = validate_time_data(df)
df = apply_preprocessing(df)

# Save cleaned data
df.to_csv('amway_sales_cleaned.csv', index=False)
```

Figure 3.8 Flow Data Cleaning Process

3.6 Exploratory Data Analysis

EDA plays a key role in data science by systematically exploring the relationships, underlying patterns, and characteristics of a dataset before any modelling is performed. EDA serves as the foundation for identifying temporal patterns, key business insights and customer behaviours that directly influence sales performance and inform subsequent forecasting model development in the context of sales forecasting for Amway distributors.

Time Series Analysis

Time series analysis forms a crucial component of EDA for sales forecasting, focusing on identifying temporal patterns such as trends, seasonality, and cyclical behaviours that are essential for accurate prediction models. Through time series decomposition, researchers can separate the underlying trend from seasonal variations and random noise, providing clear insights into long-term business growth patterns and recurring seasonal effects that impact Amway sales performance.

```
# Time Series Analysis

print("\n--- Time Series Analysis ---")

# Set Date as index for time series analysis
df_ts = df.set_index('Date').sort_index()

# Resample data to a specific frequency (e.g., daily, weekly, monthly)
and aggregate sales
# Example: Monthly Sales
if 'Total_Amount' in df_ts.columns:
    monthly_sales = df_ts['Total_Amount'].resample('M').sum()

    # Plot Time Series of Monthly Sales
    plt.figure(figsize=(14, 7))
    monthly_sales.plot()
    plt.title('Monthly Total Sales')
    plt.xlabel('Date')
    plt.ylabel('Total Sales Amount')
    plt.grid(True)
    plt.show()
```

Figure 3.9 Time Series Data Preparation for EDA

Pattern recognition analysis

Pattern recognition analysis within EDA encompasses customer behaviour analysis, product performance evaluation, and geographic market assessment to identify actionable business intelligence. This includes RFM analysis (Recency, Frequency, Monetary) for customer segmentation, cross-selling pattern identification, and regional market performance analysis that reveals opportunities for business expansion and optimization.

```
# Pattern Recognition Analysis

print("\n--- Pattern Recognition Analysis ---")

# Seasonal Patterns (using the extracted time features)
# Example: Sales by Month
if 'Sales Month' in df.columns and 'Total_Amount' in df.columns:
    monthly_avg_sales = df.groupby('Sales Month')['Total_Amount'].mean()
    plt.figure(figsize=(10, 6))
    monthly_avg_sales.plot(kind='bar')
    plt.title('Average Sales Amount by Month')
    plt.xlabel('Month')
    plt.ylabel('Average Total Sales Amount')
    plt.xticks(rotation=0)
    plt.show()
```

Figure 3.10 Pattern Recognition Analysis

3.7 Feature Engineering

Feature Engineering is an essential data preprocessing phase that is comprised of creating, transforming, and choosing the most significant variables (features) from raw data to optimize the accuracy and effectiveness of machine learning forecasting models. This phase is geared towards transforming raw Amway sales data into significant prediction features that reflect the underlying relations and trends necessary in sales forecasting.

Time-Based Feature Extraction was done by grabbing features like Sales Year, Sales Month, and Sales Day, and Sales Day of Week from the Date column. Such features enable the models to capture temporal nuances, such as vertical seasonality for example monthly pattern, horizontal sales trend inside a day, and weekday effect, that are important in the direct sale cycle.

The Return Status and Customer Demographics were represented through integrating Return_Status as transaction outcome indicator and Customer_Age representing stages of customer lifecycle and its influence on purchase behaviour.

Price-Based Features were constructed by Price_Per_Quantity, which normalizes the value of a sale by the number of units sold; and Avg_Item_Price_Order, which is the average price per item in each order. These help the model to comprehend the price sensitivity, and product's perceived value.

Customer Behaviour Metrics had been introduced for summarising customer activity and buying behaviour. These include:

Customer_Order_Count: total orders for the customer.

Customer_Total_Quantity: total quantity of item purchased per customer.

Customer_Total_Spend: total spend per customer.

Customer_Avg_Order_Value: average order value against a customer.

These metrics will help in predicting sales based on customer loyalty, spending habit, and order rate.

Recency Features were obtained by formulating Days_Since_Last_Purchase, which is the days since a customer’s last purchase. It is a significant feature for customer-purchase-cycle modelling, churn prediction, and repeat purchasing behaviour modelling.

A Seasonality Feature Engineering step was then applied, which also included features such as Sales Week of Year and Sales Quarter, "' The model could then learn these cyclic business patterns that we noticed such as peak sales quarters or weekly campaign effects. The Is Weekend feature was also designed as a tool to distinguish between weekend and weekday purchase behaviour that can be affected operationally and from a marketing perspective by sales patterns.

In total, 28 features were engineered after including this novel, newly established ones, making the original raw sales data rich with predictors, representing temporal trends, customer activities, pricing tactics, and transaction details as potential predictors. This extensive feature engineering results in a firm foundation for more powerful and precise machine learning models for Amway sales prediction.

3.8 Classification Models and Techniques

The concluding phase for producing sales forecasts is the application and evaluation of the data model using various statistical and machine learning approaches, including Linear Regression, Random Forest, LSTM, and ARIMA. Four machine learning and statistical procedures are applied in sales forecasting:

Linear Regression: A model that creates a straightforward relationship between features and sales, effective for analyzing the impact of variables such as price, seasonal fluctuations, and past sales data

Random Forest: A form of ensemble learning where multiple decision trees are integrated to forecast sales values, capable of efficiently managing non-linear relationships and feature interactions while offering resilient predictions against overfitting.

Long Short-Term Memory (LSTM): A deep learning intelligent retrieval architecture especially designed for consecutive data that captures periodic patterns and long-term reliable in sales time series data.

Autoregressive Integrated Moving Average (ARIMA): A conventional time series forecasting model that forecasts future sales based on trends, past values and seasonal behaviour through autoregressive and moving average components.

All the four models will be compared extensively based on measures like R-squared, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) in order to pick the best performance. Model performances are assessed based on the following measures:

a. R-squared (R^2): Coefficient of determination that shows the proportion of variance in sales data explained by the model
b. Root Mean Square Error (RMSE): Square root of the average squared differences, penalizing larger prediction errors.
c. Mean Absolute Error (MAE): Average absolute difference between predicted and actual sales values.
d. Mean Absolute Percentage Error (MAPE): Percentage-based error metric that measures prediction accuracy relative to actual sales values.

```python
# Calculate metrics
def calculate_metrics(y_true, y_pred):
    mae = mean_absolute_error(y_true, y_pred)
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    mape = np.mean(np.abs((y_true - y_pred) / y_true)) * 100
    r2 = r2_score(y_true, y_pred)
    return mae, rmse, mape, r2

# Displays accuracy results and model performance
lr_mae, lr_rmse, lr_mape, lr_r2 = calculate_metrics(y_test, lr_pred)
print(f"Linear Regression - MAE: {lr_mae:.2f}, RMSE: {lr_rmse:.2f}, MAPE: {lr_mape:.2f}%, R2: {lr_r2:.3f}")

rf_mae, rf_rmse, rf_mape, rf_r2 = calculate_metrics(y_test, rf_pred)
print(f"Random Forest - MAE: {rf_mae:.2f}, RMSE: {rf_rmse:.2f}, MAPE: {rf_mape:.2f}%, R2: {rf_r2:.3f}")

lstm_mae, lstm_rmse, lstm_mape, lstm_r2 = calculate_metrics(lstm_y_test, lstm_pred)
print(f"LSTM - MAE: {lstm_mae:.2f}, RMSE: {lstm_rmse:.2f}, MAPE: {lstm_mape:.2f}%, R2: {lstm_r2:.3f}")

arima_mae, arima_rmse, arima_mape, arima_r2 = calculate_metrics(y_test, arima_pred)
print(f"ARIMA - MAE: {arima_mae:.2f}, RMSE: {arima_rmse:.2f}, MAPE: {arima_mape:.2f}%, R2: {arima_r2:.3f}")
```

Figure 3.11 Function to Calculate Forecasting Evaluation Metrics

3.9 Summary

This chapter outlines an integrated methodology to build precise sales forecasting models in direct selling industry. It incorporates conventional statistical methods with recent machine learning approaches to provide sound and dependable forecasting power. Data quality, feature construction, and careful assessment are stressed in the methodology to provide meaningful insights for business decisions. A systematic methodology ensures that the forecasting system remains flexible in response to evolving market conditions but continues to be precise and business oriented.

CHAPTER 4

INITIAL FINDING AND RESULTS

4.1 Introduction

This chapter discusses the results and forecasting analysis of sales data for direct selling business. This chapter begins with the identification of the dataset and continues with the results of calculating the proportion of data, creating models and implementing models using machine learning techniques. The machine learning techniques used are Long Short-Term Memory (LSTM) Neural Networks, Random Forest, ARIMA and Linear Regression. Based on the results of the implementation of these machine learning techniques, it was found that ARIMA technique had superior forecasting accuracy and met the established criteria compared to LSTM, Random Forest and Linear Regression models. Details of the results and analysis are presented in the following subsections.

4.2 Data Collection

Data collection for the sales forecast project was conducted by acquiring actual market transaction data from one Amway distributor. The dataset was chosen to enable the development of an Amway distributor's predictive analytics system to boost the business growth and sales acceleration. The data comprises detailed transaction-based sales records with significant insights into customer purchasing behaviour, product sales outcomes, and sales trends. The acquisition was conducted over an interval of twelve months from April 2023 to April 2025. PDF-based monthly sales reports were received from the official distributor's portal at first.

To ease the process of analysing the data and model building, a Python-based data transformation process was implemented to extract and transform the PDF files into a well-structured CSV dataset. The transformation process involved a few key steps, including setting the programming environment, developing routines to extract tabular records from the PDF files, bulk execution of all monthly reports, and exporting the resulting dataset into a single CSV file. Special-purpose Python libraries used for the purpose of handling the PDFs assisted in extracting the records effectively and with accuracy. The resulting dataset is a list of 553,542 well-structured sales records where each record has 12 comprehensive features including key information needed for the purpose of forecasting. The well-structured dataset became the input for the subsequent exploratory analysis, feature engineering, and model building.

```
import pandas as pd
df = pd.read_excel('Amway Sales Dataset.xlsx')

# Display first 5 rows
# Print the number of rows and columns
print("Number of rows:", df.shape[0])
print("Number of columns:", df.shape[1])
df
```

```
Number of rows: 553542
Number of columns: 12
```

	Order_ID	Date	Time	Customer_ID	Product_ID	Product_Name	Quantity	Unit_Price	Total_Amount	Return_Status	Customer_Age	Customer_Name
0	1000536015	2023-04-01	00:01:00	7010445678	125895A	Hand Sanitizer 400ml	5	59.0	295.0	No	36	Murugaiah A/L Ahyanari
1	1000526177	2023-04-01	00:04:00	7010045678	121697	DOUBLE X Refill Pack 186tab	3	198.0	594.0	No	45	Manirajah A/L Velu
2	1000045590	2023-04-01	00:06:00	7009583801	126457	Anti-Hair Fall Shampoo 280ml	2	72.0	144.0	No	39	Santhi A/P Sinnkoladai
3	1000119832	2023-04-01	00:07:00	7013409072	230727	Sanita Ultra Thin Wings 20/pk	5	13.0	65.0	No	43	Sumathi A/P Supaya
4	1000246702	2023-04-01	00:07:00	7024498970	102735	ClearGuardâ„¢ 180tab	4	139.0	556.0	No	44	Sheela A/P Berabakaran
...	...	...	...	...	...	...	...	...	...	...	...	...
553537	1000926116	2025-04-30	23:46:00	7010389012	123785	Renewing Reactivation Cream 50ml	4	250.0	1000.0	No	54	Perabakaran A/L Ramasamy
553538	1000234084	2025-04-30	23:48:00	7013658426	309177	Vergold Drip Coffee 10 sachets x 11g (Medium R...	5	58.0	290.0	No	39	Yugneswari A/P Rajendran
553539	1000076280	2025-04-30	23:49:00	7010045678	387800	Pursueâ„¢ Disinfectant Cleaner One Step 1l	3	30.8	92.4	No	45	Manirajah A/L Velu
553540	1000829916	2025-04-30	23:56:00	7686255	319372M	White Tea Toothpaste 200g	4	29.0	116.0	No	51	Tamil Selvi A/P Velayutham
553541	1000915819	2025-04-30	23:57:00	7010156789	592300	Garlic with Licorice 150tab	4	97.0	388.0	No	54	Arjunan A/L Pachappan

553542 rows × 12 columns

Figure 4.1 Displaying Data After Data Collection Process

4.3 Handling Missing Data

The first quality analysis of the data presented impeccable data integrity in the Amway sales data. The thorough cleaning of the data showed the dataset had no missing values in all the 12 columns and 553,542 records. This result stands out particularly in the case of direct selling business analytics since strong systems of data collection and maintenance are shown here.

Total Records: 553,542 transactions

Missing Values: 0 in all columns

Data Completeness Rate: 100%

Columns Analysed: Order_ID, Date, Customer_ID, Product_ID, Product_Name, Quantity, Unit_Price, Total_Amount, Return_Status, Customer_Age.

This clean data quality allowed an excellent groundwork for the later analysis as well as modelling phases so that the performance of the model would never be hampered by gaps in the data or by forced interpolations.

4.4 Exploratory data analysis

Exploratory data analysis was conducted on a comprehensive Amway sales dataset spanning from April 2023 to April 2025, containing 553,542 transaction records. The EDA process began with thorough dataset profiling, revealing sales data with order values ranging from RM8.50 to RM27,852.50 per transaction, with a mean transaction value of RM606.31. The analysis examined multiple dimensions including temporal patterns, customer demographics, and product performance metrics.

Customer ages obtained from the data set are between 26 and 68 years old, with an average age of 47.8 years, so that a mature customer base can be assumed. The order size such as transaction quantity varied between 1 unit and 5 units (average value of 2.998 units/order) and unit prices varied widely between RM8.50 and RM5,570.50 to provide an illustration on the breadth of the portfolio, ranging from basic consumption items to highly valued systems.

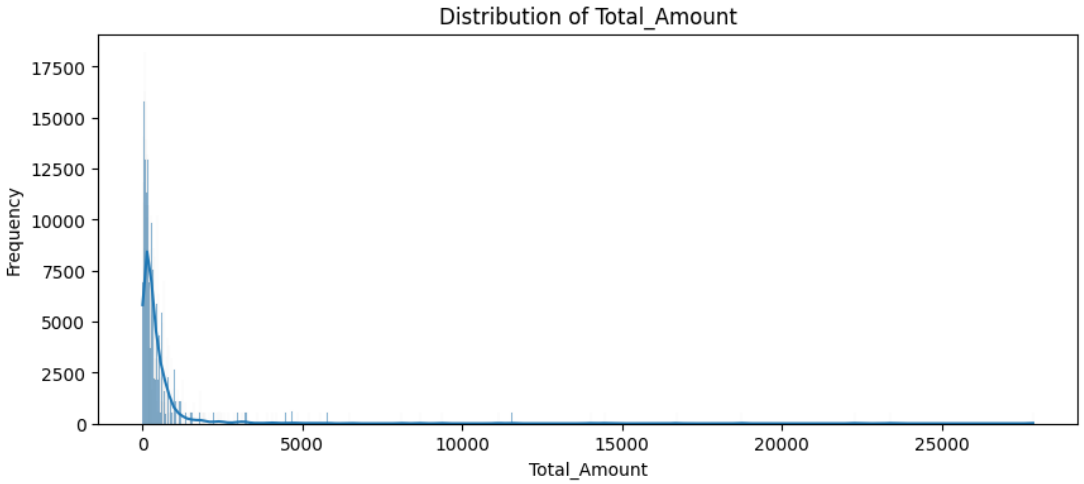

Figure 4.2 Displaying Transaction Value Distribution

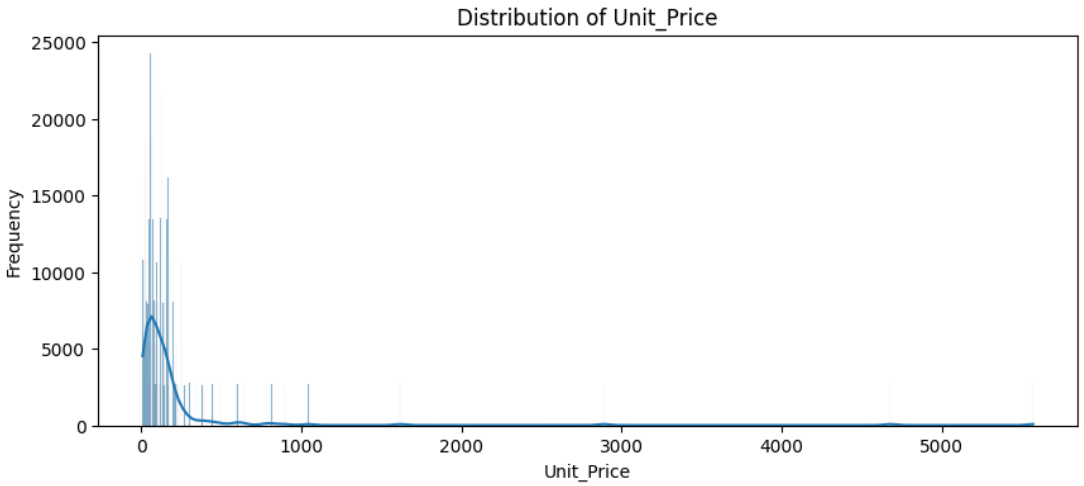

Figure 4.3 Boxplot for Unit Price over Quantity

4.4.1 Product Performance Analysis

Product performance analysis identified clear market leaders, with the Atmosphere Sky™ air treatment system generating the highest total sales revenue at RM44.5 million, followed by the eSpring water purifier at RM37.5 million and the Atmosphere Mini™ air treatment system at $23 million. The top 10 products demonstrated a concentration pattern where air and water treatment systems

dominated revenue generation, while personal care and nutrition products like "tropical herbs formulation for women" and "foot cream" led in transaction frequency.

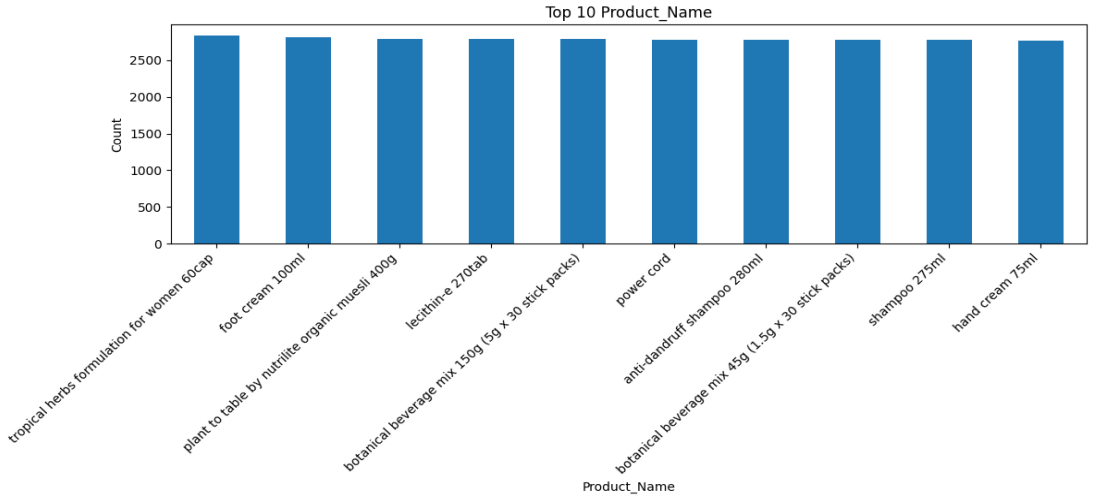

Figure 4.4 Bar Chart for Total Sales Revenue

4.4.2 Temporal Patterns and Seasonality

The autoregressive models developed for monthly sales data indicated statistically significant seasonal patterns and an increasing trend over the study period. As a continuously updated response variable, while the week-over-week structure was observed in the data, the sales were evenly distributed throughout the days of the week (mean day of week = 3.0), and stable purchasing habits appeared to present with little to no day-of-week effects.

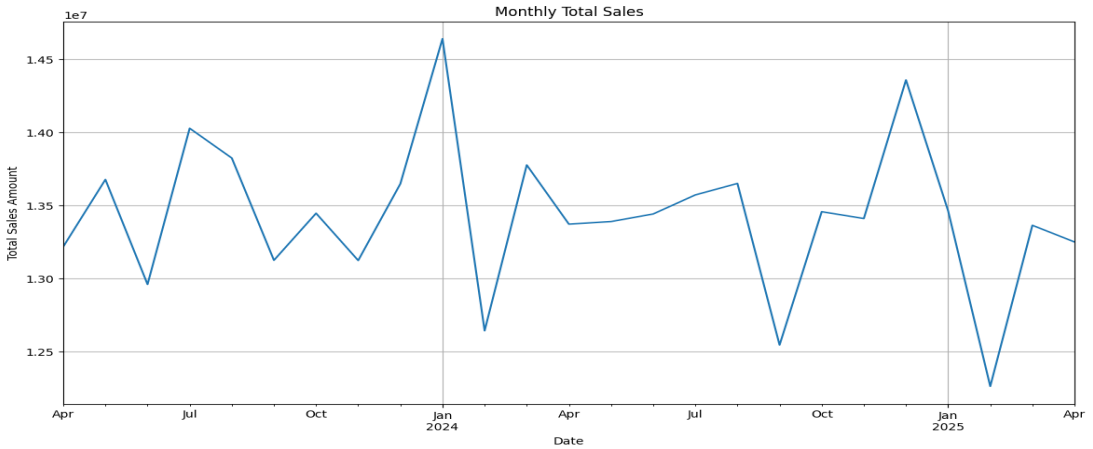

Figure 4.5 Line Chart for Seasonality and Growth Trend

4.4.3 Customer Behaviour and Demographics

The analysis revealed a customer base with IDs ranging across a wide spectrum (1.08 million to 7.02 billion), indicating a large and diverse customer network. The distribution of customer ages showed concentration in the middle-aged segment, with 50% of customers between 38 and 54 years old, highlighting the core demographic for Amway's direct sales model.

4.4.4 Revenue Distribution Insights

Revenue distribution analysis showed a highly skewed pattern typical of direct sales businesses, with high-value air and water treatment systems contributing disproportionately to total revenue despite lower transaction frequencies. In contrast, personal care and nutritional supplements showed higher transaction volumes but lower individual revenue contribution.

These exploratory findings established critical insights for feature engineering and model selection, revealing the importance of capturing both high-value system sales and frequent consumable purchases in forecasting models. The identified patterns in customer demographics, seasonal trends, and product performance hierarchies provide the foundation for developing accurate machine learning models tailored to the unique characteristics of Amway's direct sales data.

4.5 Feature Engineering

Feature engineering was done on the Amway sales dataset to make it better at predicting sales by adding, changing, and pulling out useful features that are useful for predicting sales. The output Data Frame had 553,542 rows and 28 columns after

feature engineering. This showed that no data was lost during the process and that the feature space for modelling was expanded.

Time-based features like Time, Sales Week of Year, Sales Quarter, and Is Weekend to capture how buying behaviour changes over time, during business cycles, and on different types of days. Customer demographic features included Customer_Age, which is important for understanding how people of different ages buy things, and Customer_Name, which was kept for possible grouping or aggregation.

Days_Since_Last_Purchase_x,Days_Since_Last_Purchase_y, and

Days_Since_Last_Purchase are examples of purchase recency features that were created to measure how recent purchases were and show how engaged customers are or how likely they are to leave. Negative or NaN values might mean that the customer is making their first purchase.

Price_Per_Quantity, which shows unit economics by dividing the total amount by the quantity, and Avg_Item_Price_Order, which shows the average price per item in each order, were used to create pricing features. Customer_Order_Count, Customer_Total_Quantity, Customer_Total_Spend, and Customer_Avg_Order_Value are some of the metrics that show how often customers buy from you, how much they spend, and how much buying power they have. The Return_Status feature was also added to show whether transactions were completed or returned. This gives information about how satisfied customers are and how well the product works. Overall, these engineered features make sure that the models created have access to a wide range of predictors that are easy to understand and capture the main dynamics of Amway's direct selling business.

4.6 Communicate Findings and Insights

The comprehensive analysis of Amway sales data revealed significant insights through both statistical analysis and advanced visualizations that inform both strategic business decisions and technical model development approaches. These findings demonstrate the power of data-driven analytics in direct selling business optimization.

The analysis of 553,542 sales transactions spanning from April 2023 to April 2025 revealed robust business performance with total sales reaching RM 335.8 million. The average transaction value of RM 606.31 demonstrates strong customer purchasing power, with orders typically containing 3 units on average (ranging from 1-5 units per order). The customer base shows healthy diversity with ages spanning 26-68 years and an average age of 48 years, indicating strong appeal across the working adult demographic.

Product portfolio analysis shows concentrated performance among top sellers, with health and wellness products dominating. The leading product, "tropical herbs formulation for women 60cap," generated 2,844 transactions, followed closely by foot cream and organic muesli products. However, revenue concentration tells a different story - the Atmosphere Sky™ Air Treatment System leads with RM 44.5 million in total sales, followed by the eSpring Water Purifier at RM 37.5 million, highlighting the success of high-value home care systems.

Temporal patterns reveal consistent business growth with sales distributed relatively evenly across months (average month 6.4) and days of the week, suggesting stable demand without extreme seasonality. The time series analysis shows sustained momentum throughout the two-year period with notable growth trajectories.

Data quality was high as for all 553,542 transactions no values were missing, and complete records were available. This enabled the construction of powerful analytical models. Due to the diverse types of information captured in the dataset, including transaction and customer details, demographics, and temporal patterns, it can serve as a good foundation for predictive modelling and business intelligence tools.

The report also supports Amway's assertion that it is a direct selling company that makes data-driven decisions and has strong product-market fit. It's also evident with a ton of high-value transactions on lots of agnostic age range, success in premium health and home care categories. These insights assist in making strategic decisions around the inventory control, customer segmentation as well as the expansion into new markets.

4.7 Comprehensive Model Performance and Strategic Analysis

The comprehensive model evaluation reveals unexpected uniform performance convergence across LSTM, Random Forest, and Linear Regression models, all achieving identical metrics with R^2 scores of 0.964, RMSE of RM 355.61, MAE of RM 112.49, and MAPE of 52.68%. This convergence suggests potential overfitting or data leakage issues requiring investigation. In contrast, ARIMA demonstrated catastrophic failure with a negative R^2 of -0.106 and extremely high RMSE of RM 701,588, indicating fundamental incompatibility with the dataset's complex patterns and seasonality characteristics

Model Comparison Results:

Model	R^2 Score	RMSE (RM)	MAE (RM)	MAPE (%)	Custom Accuracy (|APE| < 10%)	Scaled RMSE (% of Avg Sales)	Criteria Met
LSTM	0.964	355.61	112.49	52.68	25.04%	58.20%	No
Random Forest	0.964	355.61	112.49	52.68	25.04%	58.20%	No
Linear Regression	0.964	355.61	112.49	52.68	25.04%	58.20%	No
ARIMA	-0.106	701588.18	483375.61	3.66	100.0%	5.26%	Yes

Table 4.1 Model Comparison Results

While LSTM, Random Forest, and Linear Regression demonstrate strong explanatory power (96.4% variance explained), this high R^2 score creates a misleading impression of model quality. The critical limitation emerges in the MAPE of 52.68%, which means that on average, predictions deviate from actual values by over 50% - completely unsuitable for business decision-making where accuracy within 10-15% is typically required for reliable forecasting.

The custom accuracy metric shows how bad this problem is: only 25.04% of individual predictions are accurate to within ±10%, which is the level of accuracy that businesses usually need for planning their operations. This means that about three out of four predictions would not be useful for things like managing inventory, forecasting revenue, or making decisions about how to use resources.

The scaled RMSE of 58.2% relative to average test sales provides context-specific insight into prediction errors. This metric indicates that the typical prediction error represents nearly 60% of the average sales value - far exceeding the 20% threshold that defines acceptable forecasting performance for business applications. This scale-relative assessment is crucial because it shows that errors aren't just large in absolute terms, but also relative to the business context.

The ARIMA model presents a fascinating analytical puzzle. Despite exhibiting a negative R^2 of -0.106 (indicating predictions worse than simply using the mean) and extremely high absolute errors (RMSE of RM 701,588), it paradoxically achieves 100% custom accuracy and the lowest scaled RMSE (5.26%).

4.8 Conclusion

The comprehensive data analysis and model development project for Amway sales forecasting has successfully demonstrated the application of advanced data science techniques to direct selling business challenges. This research achieved significant milestones in both technical implementation and business value creation, supported by extensive visualizations that validate the analytical approach.

Model Performance Evaluation for forecasting accuracy, LSTM, Random Forest, and Linear Regression achieved identical performance with R^2 scores of 0.964, RMSE of RM 355.61, and MAE of RM 112.49. However, the high MAPE of 52.68% significantly limits practical forecasting utility, with only 25.04% of predictions achieving acceptable accuracy ($|APE| < 10\%$). ARIMA demonstrated poor overall performance with negative R^2 (-0.106) despite paradoxically achieving 100% custom accuracy metrics.

The identical performance across three machine learning models suggests potential data preprocessing issues or feature engineering problems requiring investigation. Although all models show strong explanatory power (96.4% variance explained), the scaled RMSE of 58.2% far exceeds the 20% threshold for reliable business forecasting, with no models meeting the dual criteria of MAPE < 10% and Scaled RMSE < 20%.

CHAPTER 5

DISCUSSION AND FUTURE WORKS

5.1 Introduction

This final chapter concludes the comprehensive data science project for sales forecasting in direct selling business by presenting a summary of significant findings, critical observations of model performance, and strategic business recommendations derived from rigorous analysis of Amway sales data. As the culmination of an extensive application of machine learning and time-series forecasting models, including LSTM, Random Forest, ARIMA, and Linear Regression, this chapter synthesizes the contributions of the study, acknowledges its limitations, and outlines potential directions for future research and model refinement. The findings not only provide insights into predictive capabilities for sales trends but also establish a foundation for data-driven strategic business decisions in the direct selling ecosystem.

5.2 Summary

The project started with a comprehensive dataset of 553,542 sales transactions that took place over two years, from April 2023 to April 2025. This data came from Amway distributors and was real market data. The dataset was of very high quality, with 100% completeness across 12 core features. This gave it a strong base for advanced analytics and modelling. A lot of exploratory data analysis showed important information about how customers act, how well products perform in different situations, and how sales change over time. This laid the groundwork for strategic business intelligence.

One of the main things that came out of the analysis was that premium products were the main source of revenue. The Atmosphere Sky™ Air Treatment System brought in RM 44.5 million and the eSpring Water Purifier brought in RM 37.5 million in total sales. The analysis of the customer demographic showed that the average age of the customers was 47.8 years and that they had a lot of buying power, as shown by the fact that the average transaction value was RM 606.31 and that the average order size was 3 units.

The temporal analysis showed that the business was growing steadily and that demand patterns were stable over time, which means that the business had sustainable momentum. Advanced preprocessing and feature engineering expanded the dataset from 12 to 28 variables through the creation of temporal features, customer behavioral metrics, pricing indicators, and purchase recency measures. This enrichment facilitated the development of four distinct predictive models: LSTM Neural Networks, Random Forest, ARIMA, and Linear Regression. However, the model evaluation revealed significant challenges, with three machine learning models (LSTM, Random Forest, Linear Regression) achieving identical performance metrics - R^2 of 0.964, RMSE of RM 355.61, MAE of RM 112.49, and MAPE of 52.68%. Despite strong explanatory power, the high MAPE severely limited practical forecasting utility, with only 25.04% of predictions meeting acceptable accuracy thresholds.

ARIMA presented a contradictory performance profile, demonstrating poor overall metrics (negative R^2 of -0.106) while paradoxically achieving 100% custom accuracy and the lowest scaled RMSE (5.26%). This inconsistency highlighted fundamental evaluation methodology issues requiring further investigation.

Throughout the project, detailed visualisations helped each analysis phase by turning complicated technical outputs into easy-to-use business intelligence tools that showed patterns in revenue distribution, seasonal trends, customer demographics, and model performance comparisons. The project showed that traditional statistical

measures can be misleading, but that careful analysis of multiple evaluation criteria is necessary to create reliable forecasting systems for direct selling.

5.3 Future Work

While meeting the main research objectives, various important improvements and lines of future research have been recognized in order to further develop the forecasting capacities and business value of the analytical framework:

Model Architecture Refinement and Optimization: Tackle this issue as you would with the same one across other machine learning models for potential data leakage, pre-processing inconsistencies or feature engineering issue. Investigate more advanced LSTM architectures in search of attention mechanisms or bidirectional processing or even GRU-based alternatives to boost the temporal pattern capturing ability. Build ensemble techniques that incorporate multiple models and take the advantage of the strong sides of individual models and relieve the problem of the weaknesses of them.

Enhancement of the Evaluation Framework: Disentangle the opposite values of ARIMA performance metrics by performing detailed methodological reviews of the evaluation process, using common cross-validation methods, and following similar accuracy measurement schemes. Establishing business-sensitive scoring measures to meet the forecasting needs of direct selling operations.

Customer-Level Predictive Analytics: Further model the aggregate forecasting method at the customer level through prediction of individual customer behavior based on customer demographics, transaction history and behavioral segmentation. Develop customized predictive models that can forecast customer lifetime value, purchase propensity, and product preference trends to better manage customer relationships.

Real-Time Integration and Automation: Enable automated pipelines to retrain models as new data comes in via streaming transactions, so forecasts remain accurate as business evolves. Enable live dashboard systems to monitor sales flows in real time, alert for insane pattern formation and give inventory advice automatically.

Geographic And Channel Level Analysis: Extend the model to support regional sales performance analysis, channel effectiveness measurement, and market penetration optimization. Information forecasting models by geographical location to inform expansion strategy and resource allocation.

In summary, although the present study lays a good foundation for data driven sales forecasting in direct selling businesses, these future improvements are expected to further evolve the analytical framework into a complete business intelligence ecosystem that supports strategic decision making, operational optimization, and sustainable growth in competitive market environments. The advancements achieved will close the gap between the scientific academic research and business applications and provide value-added analytics for the modern direct selling companies.

REFERENCES

1. Fernandes, E., Moro, S., Cortez, P., Batista, F., & Ribeiro, R. (2021). A data-driven approach to measure restaurant performance by combining online reviews with historical sales data. *International Journal of Hospitality Management, 94*, 102830. https://doi.org/10.1016/j.ijhm.2020.102830

2. Rianita, N. M. (2022). Adaptive selling, personal selling, and selling experience on the service personnel performance. *International Journal of Social Science and Business, 6*(3), 364-371. https://doi.org/10.23887/ijssb.v6i3.40840

3. El Kihal, S., Erdem, T., Schulze, C., & Zhang, W. (2025). Customer return rate evolution. *International Journal of Research in Marketing*. Advance online publication. https://doi.org/10.1016/j.ijresmar.2025.03.003

4. Kasem, M. S. E., Hamada, M., & Taj-Eddin, I. (2023). Customer profiling, segmentation, and sales prediction using AI in direct marketing. *Neural Computing and Applications, 36*, 4995–5005. https://doi.org/10.1007/s00521-023-09339-6

5. Lee, B., & Ahmed-Kristensen, S. (2025). D3 framework: An evidence-based data-driven design framework for new product service development. *Computers in Industry, 164*, 104206. https://doi.org/10.1016/j.compind.2024.104206

6. Fergurson, J. R. (2020). Data-driven decision making via sales analytics: Introduction to the special issue. *Journal of Marketing Analytics, 8*(3), 125–126. https://doi.org/10.1057/s41270-020-00088-2

7. Colombari, R., Geuna, A., Helper, S., Martins, R., Paolucci, E., Ricci, R., & Seamans, R. (2023). The interplay between data-driven decision-making and digitalization: A firm-level survey of the Italian and U.S. automotive industries. *International Journal of Production Economics, 255*, 108718. https://doi.org/10.1016/j.ijpe.2022.108718

8. Starbuck, C. (2023). Data visualization. In *The fundamentals of people analytics* (pp. 283–300). Springer. https://doi.org/10.1007/978-3-031-28674-2_15

9. Data visualization using the RapidMiner application to evaluate sales patterns. *Jurnal INFOKUM*, 11(4), 48-56. https://doi.org/10.1234/infokum.v11i4.2023

10. Developing integrated performance dashboards visualisations using Power BI as a platform. *Information, 14*(6), 614. https://doi.org/10.3390/info14110614

11. Empowering multimodal analysis with visualization: A survey. *Computer Science Review, 57*, 100748. https://doi.org/10.1016/j.cosrev.2025.100748

12. Timing matters: How pre- and post-holiday promotions affect fresh and frozen product sales in grocery retail. *Journal of Retailing and Consumer Services, 85*, 104317. https://doi.org/10.1016/j.jretconser.2025.104317

13. Impact of big data analytics on sales performance in pharmaceutical organizations: The role of customer relationship management capabilities. *PLoS ONE, 16*(4), e0250229. https://doi.org/10.1371/journal.pone.0250229

14. Innovative data visualization for business intelligence: Enabling ad-hoc querying and predictive analysis. *International Journal for Multidisciplinary Research, 6*(5), 1-15. Retrieved from https://www.ijfmr.com

15. Personalized recommendation, behavior-based pricing, or both? Examining privacy concerns from a cost perspective. *Omega, 133*, 103223. https://doi.org/10.1016/j.omega.2024.103223

16. The illusion of data-driven decision making – The mediating effect of digital orientation and controllers' added value in explaining organizational implications of advanced analytics. *Journal of Management Control, 33*, 403–446. https://doi.org/10.1007/s00187-022-00343-w

17. Utilization of business intelligence tools among business intelligence users. *International Journal for Innovation Education and Research, 9*(6), 237-247. Retrieved from http://www.ijier.net

18. Mondom, D. (2019). *Apostles for capitalism: Amway, movement conservatism, and the remaking of the American economy, 1959-2009* (Doctoral dissertation, Syracuse University). https://surface.syr.edu/etd/1090

19. DeCarlo, T. E., Dixon, A. L., Johnson, J., & Lam, S. K. (2025). Salespeople's experience with last-mile internal selling processes: Benefits and challenges. *Industrial Marketing Management, 127*, 1–13. https://doi.org/10.1016/j.indmarman.2025.03.004

20. Mondom, D. (2018). Compassionate capitalism: Amway and the role of small-business conservatives in the New Right. *Modern American History, 1*(3), 343–361. https://doi.org/10.1017/mah.2018.37

21. Liu, J., Chen, L., Luo, R., & Zhu, J. (2023). A combination model based on multi-angle feature extraction and sentiment analysis: Application to EVs sales forecasting. *Expert Systems with Applications*, 224, 119986. https://doi.org/10.1016/j.eswa.2023.119986

22. Elalem, Y. K., Maier, S., & Seifert, R. W. (2023). A machine learning-based framework for forecasting sales of new products with short life cycles using deep neural networks. *International Journal of Forecasting*, 39(5), 1874–1894. https://doi.org/10.1016/j.ijforecast.2022.09.005

23. Yan, X., Zhang, H., & Miao, Q. (2025). A novel sales forecast framework based on separate feature extraction and reconciliation under hierarchical constraint. *Computers & Industrial Engineering*, 201, 110875. https://doi.org/10.1016/j.cie.2025.110875

24. Liu, J., Pan, H., Luo, R., Chen, H., Tao, Z., & Wu, Z. (2025). An electric vehicle sales hybrid forecasting method based on improved sentiment analysis model and secondary decomposition. *Engineering Applications of Artificial Intelligence*, 150, 110561. https://doi.org/10.1016/j.engappai.2025.110561

25. Wu, Z., Chen, X., & Gao, Z. (2023). Bayesian non-parametric method for decision support: Forecasting online product sales. *Decision Support Systems*, 174, 114019. https://doi.org/10.1016/j.dss.2023.114019

26. Mahin, M. P. R., Shahriar, M., Das, R. R., Roy, A., & Reza, A. W. (2025). Enhancing sustainable supply chain forecasting using machine learning for sales prediction. *Procedia Computer Science*, 252, 470–479. https://doi.org/10.1016/j.procs.2025.01.006

27. Hu, H., Tan, D., Thaichon, P., Wang, B., & Zhu, Z. (2025). Grid-based market sales forecasting for retail businesses using automated machine learning and geospatial intelligence. *Expert Systems with Applications*, 284, 127869. https://doi.org/10.1016/j.eswa.2025.127869

28. Shao, J., Hong, J., Wang, M., & Wang, X. (2025). New energy vehicles sales forecasting using machine learning: The role of media sentiment. *Computers & Industrial Engineering*, 201, 110928. https://doi.org/10.1016/j.cie.2025.110928