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

SIVARAJAN A/L S.ESVARAN

UNIVERSITI TEKNOLOGI MALAYSIA

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.

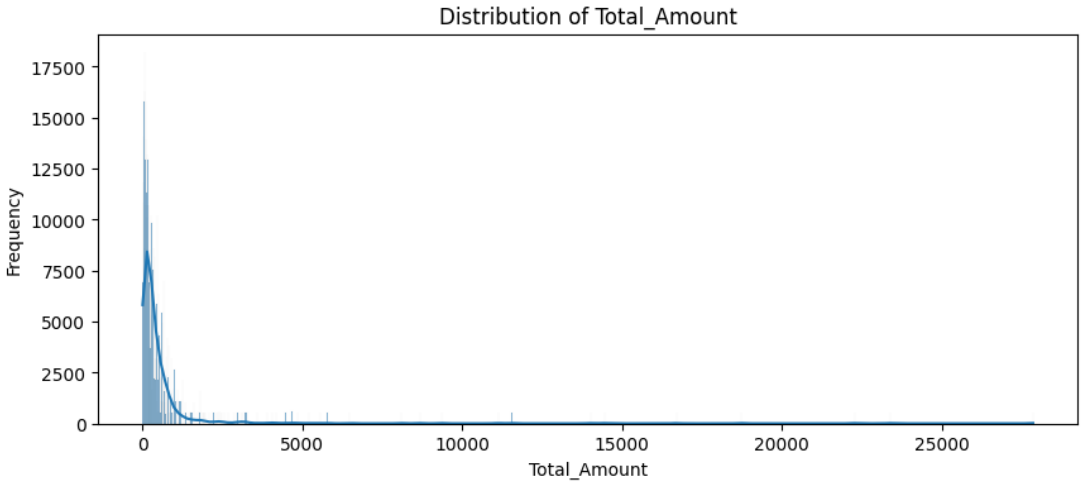

Figure 4.2 Displaying Transaction Value Distribution

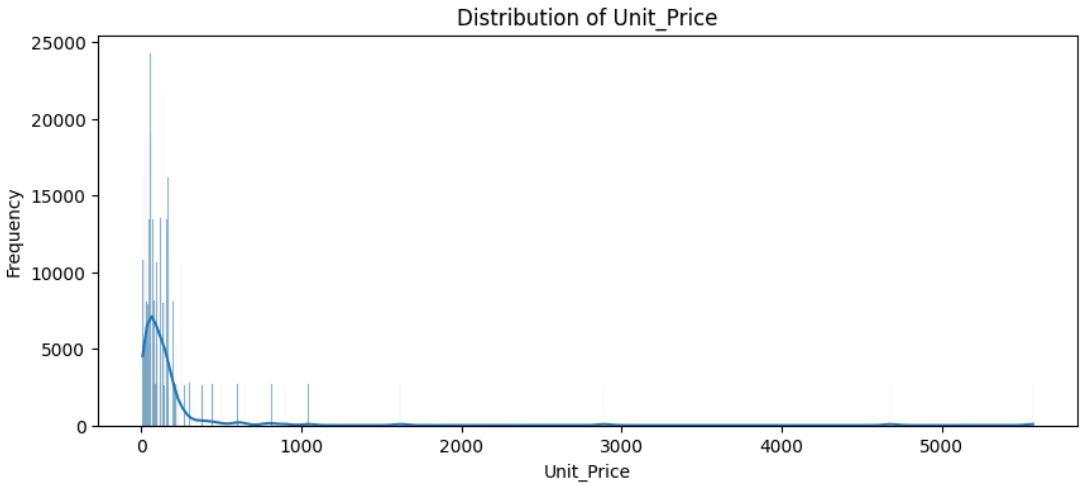

Figure 4.3 Boxplot for Unit Price over Quantity

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.

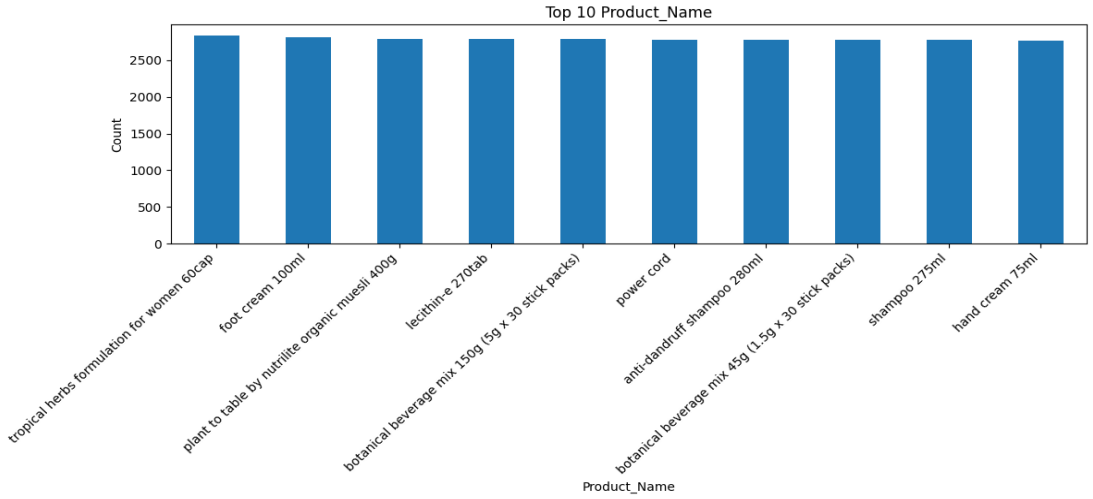

Figure 4.4 Bar Chart for Total Sales Revenue

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.

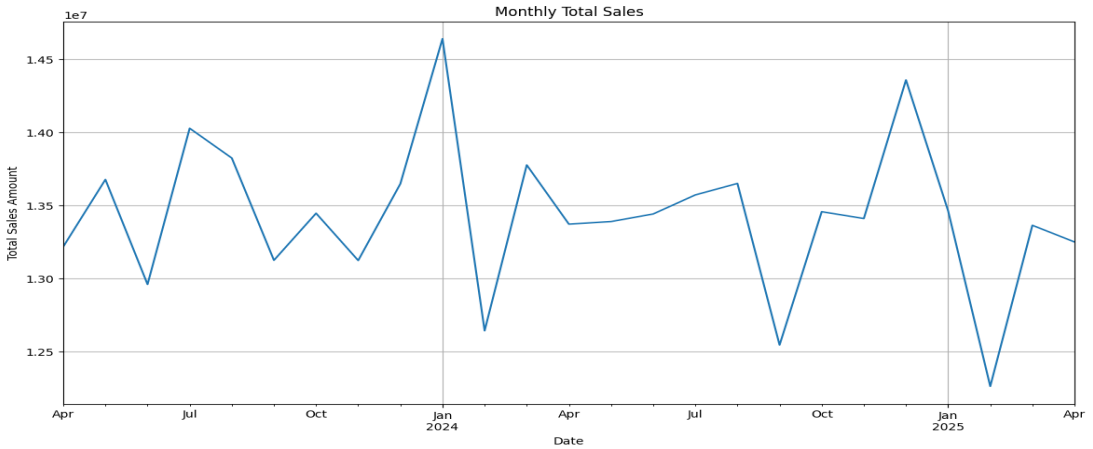

Figure 4.5 Line Chart for Seasonality and Growth Trend

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%.