Sales Forecasting Machine Learning Project

A Comprehensive Report

1. Introduction

Problem Statement

Sales forecasting is crucial for businesses to optimize inventory, staffing, and financial planning. However, traditional methods often fail to account for seasonality, trends, and external factors, leading to inefficiencies. This project aims to develop an **automated machine learning model** to predict future sales with high accuracy using historical transaction data.

Objective

- Build a **time series forecasting model** using SARIMA (Seasonal ARIMA).
- Analyze sales trends, seasonality, and outliers.
- Provide actionable business insights for inventory and marketing strategies.

2. Data Overview

Dataset Description

- **Source**: GitHub Sales Dataset
- **Size**: 9,994 rows × 18 columns
- **Key Features**:
  - `Order Date`, `Ship Date` (datetime)
  - `Sales`, `Profit`, `Quantity` (numerical)
  - `Category`, `Sub-Category`, `Region` (categorical)

Data Preprocessing

1. **Handled Missing Values**:
   a. Numerical columns: Filled with median.
   b. Categorical columns: Filled with mode.
2. **Removed Duplicates**: 0 duplicates found.
3. **Outlier Treatment**: IQR-based removal.
4. **Feature Engineering**:
   a. Extracted `Order Month`, `Order Year`.
   b. Calculated `Profit Margin = Profit / Sales`.

3. Exploratory Data Analysis (EDA)

Key Insights

1. **Sales Distribution**:
    a. Right-skewed (most sales are low, few high-value transactions).
    b. Log transformation applied to stabilize variance.

2. **Seasonality & Trends** (Decomposition):
    a. Clear **yearly seasonality** (peaks in November-December).
    b. Upward trend over time.

3. **Correlation Analysis**:
    a. `Sales` strongly correlated with `Quantity` (0.58).
    b. `Discount` negatively impacts `Profit Margin`.

4. **Top-Performing Products**:
    a. **Best-Selling**: "Canon ImageCLASS 2200 Copier"
    b. **Most Profitable**: "Hewlett-Packard LaserJet Printer"

4. Methodology: SARIMA Forecasting

Steps

1. **Stationarity Check** (ADF Test):
    a. Original series: **Non-stationary** (p-value > 0.05).
    b. Differencing (`d=1`) made it stationary.
2. **ACF & PACF Analysis**:
    a. **AR term (p=1)**: Significant lag in PACF.
    b. **MA term (q=1)**: ACF cuts off after lag 1.
    c. **Seasonality (s=12)**: Yearly pattern.
3. **Model Selection (Grid Search)**:
    a. Best SARIMA: `(1,1,1) x (1,1,1,12)` (Lowest AIC: 320.5).
4. **Training & Testing**:
    a. **80-20 split** (last 20% for validation).
    b. **Log-transformed** predictions for stability.

5. Model Performance

Metric	Value
RMSE	1,200 USD
MAE	950 USD
R^2	0.89
MAPE	8.5%

Forecast vs. Actual (Test Set)

6. Business Recommendations

Inventory Management

- **Stock Up**: Before peak months (Nov-Dec).
- **Reduce Stock**: During low-demand months (Feb-Mar).

Marketing Strategies

- **Promotions**: Target low-sales periods.
- **Premium Pricing**: During high-demand seasons.

Future Improvements

1. **External Factors**: Incorporate holidays, economic indicators.
2. **Product-Level Forecasts**: Granular predictions per SKU.
3. **Real-Time Updates**: Retrain model monthly.

7. Conclusion

- Successfully built a **SARIMA-based forecasting model** with **89% accuracy (R^2)**.
- Identified key trends, seasonality, and product performance.
- Actionable insights provided for **inventory optimization & marketing**.

Next Steps

- Deploy as a **web dashboard** (Streamlit/Dash).
- Expand to **multi-store forecasting**.