Task instruction

Dataset Introduction

This is a transactional data set which contains all the transactions occurring between 01/12/2009 and 09/12/2011 for a non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Dataset source

Dataset link: https://archive.ics.uci.edu/static/public/502/online%2Bretail%2Bii.zip

Variable descriptions

Variable Name	Role	Type	Description	Units	Missing Values
InvoiceNo	ID	Categorical	a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation		no
StockCode	ID	Categorical	a 5-digit integral number uniquely assigned to each distinct product		no
Description	Feature	Categorical	product name		no
Quantity	Feature	Integer	the quantities of each product (item) per transaction (if quantity is negative, it means that order is cancelled)		no
InvoiceDate	Feature	Date	the day and time when each transaction was generated		no
UnitPrice	Feature	Continuous	product price per unit	sterling	no
CustomerID	Feature	Categorical	a 5-digit integral number uniquely assigned to each customer		no
Country	Feature	Categorical	the name of the country where each customer resides		no

Main task instruction

You must perform the following task on given dataset.

1. Data integration and cleaning
    a. Handle missing values, duplicates, invalid quantities/prices, and ensure date parsing is correct. The dataset contains 2 sheets.
2. Exploratory Data Analysis
    a. Include at least 3 meaningful visualizations and highlight at least 2 insights about customers/products.
3. Customer Churn Prediction:
    a. Churn = no purchase within 90 days of last purchase
    b. Use 2 years of online retail sales data to predict whether a customer will purchase again within the next 90 days after their last transaction.
    c. Train at least one classifier and evaluate the performance.
4. Sales Forecast
    a. Select top 20 products by total revenue over 2 years.
    b. Aggregate data monthly. Forecast next 3 months' sales per product (top 20) using baseline (moving average/naïve) or ML/statistical model (ARIMA, Prophet, or LightGBM). Evaluate with SMAPE.

Results should include the followings:

1. Analysis Report (can be .ppt/.pdf)
2. Analysis code(s)/notebook(s)
3. Outputs (models, predicted results, etc,..)