Product Popularity Prediction

TEAM C-7

G PRAJWAL PRIYADARSHAN(24214)

KABILAN K (24224)

KISHORE B(24227)

RAHUL L S(24249)

Introduction

- **Product Popularity Prediction** refers to the process of **predicting which products are likely to be popular** (i.e., in high demand or purchased frequently) based on **historical data** like sales, customer behavior, and product characteristics.

Introduction & Objective :

- **Objective:**
  To identify popular products by analyzing purchase behavior using machine learning.
- **Dataset:**
  Unlabeled UK-based retail data with product details, quantities sold, invoice dates, unit prices, and customer IDs.
- **Preprocessing:**
  - Removed duplicates
  - Handled missing values
  - Aggregated data by product to compute total quantity sold, revenue, and average unit price
- **Labeling:**
  Products with quantity sold above the median are labeled as ‘popular’.
- **Modeling:**
  Built Random Forest, Decision Tree, and Logistic Regression models to classify products based on popularity.

About Our Dataset

01 This is a transactional data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

02 No of Features : 8
No of Instances : 541909

Feature	Description
InvoiceNo	6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation
StockCode	a 5-digit integral number uniquely assigned to each distinct product
Description	product name
Quantity	the quantities of each product (item) per transaction
InvoiceDate	the day and time when each transaction was generated
UnitPrice	product price per unit
CustomerID	a 5-digit integral number uniquely assigned to each customer
Country	the name of the country where each customer resides

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HE	6	01-12-2010 08:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANT	6	01-12-2010 08:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEAR	8	01-12-2010 08:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLA	6	01-12-2010 08:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTI	6	01-12-2010 08:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NE	2	01-12-2010 08:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTE	6	01-12-2010 08:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UN	6	01-12-2010 08:28	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED	6	01-12-2010 08:28	1.85	17850	United Kingdom
536367	84879	ASSORTED COLOUR	32	01-12-2010 08:34	1.69	13047	United Kingdom
536367	22745	POPPY'S PLAYHOUSE	6	01-12-2010 08:34	2.1	13047	United Kingdom
536367	22748	POPPY'S PLAYHOUSE	6	01-12-2010 08:34	2.1	13047	United Kingdom
536367	22749	FELTCRAFT PRINCES	8	01-12-2010 08:34	3.75	13047	United Kingdom
536367	22310	IVORY KNITTED MU	6	01-12-2010 08:34	1.65	13047	United Kingdom
536367	84969	BOX OF 6 ASSORTED	6	01-12-2010 08:34	4.25	13047	United Kingdom
536367	22623	BOX OF VINTAGE JIG	3	01-12-2010 08:34	4.95	13047	United Kingdom

OverView

Project Goal: Predicting Product Popularity

This project uses **transactional data** from the **Online Retail dataset** to **predict product popularity** with **Machine Learning**. In today's competitive retail world, knowing what customers prefer and predicting demand is crucial. By identifying popular products early, retailers can improve **customer satisfaction**, avoid **overstocking or stockouts**, and **increase profits** through better personalized marketing and inventory management.

Methodology

This project uses transactional retail data to predict product popularity through a machine learning pipeline. After cleaning the data and removing invalid entries, key features such as total quantity sold, revenue, and invoice count are aggregated at the product level. Products are labeled as ‘popular’ or ‘not popular’ using a median-based threshold. Classification models like Random Forest are then trained on this labeled data, and performance is evaluated using accuracy, precision, recall, and F1-score. Important features influencing popularity are identified. Optionally, a simple UI (e.g., Streamlit) can be added to let users upload data and view predictions, supporting better business decisions on stocking and marketing.

Thank you!