2248 – INSY - 5377

Web And Social Analytics

Fall 2024

Professor: Dr. Riyaz Sikora

PROJECT PROPOSAL

Group 6

Devansh Pandey	1002182366
Greeshma Gajula	1002169575
Kaustubh Ashok Gawande	1002157979
Vishwaksen Reddy Manda	1002161923
Yuvarekha Mahendran	1002175347

Project Title: E-Commerce (Walmart) Sales Dataset

Data Source: Kaggle.com

URL - https://www.kaggle.com/datasets/devarajv88/walmart-sales-dataset

Project Overview:

The objective of this project is a deep comprehension of customer purchasing behavior in the case of Walmart. It is an attempt to define the patterns and trends in relation to purchase decision making by various groups of customers and the use of machine learning in the prediction of future purchases.

Given the current level of market competition especially in the retail space, it is important to determine customer preferences, the purchasing patterns and what influences high value sales. This is important in designing appropriate marketing mix, improving customer satisfaction and managing stock levels. Walmart, the world's number one retailer has vast information on customer profiles, transaction data and the products purchased by the customers. It is our objective that by doing so, we will have information that can be beneficial in the making of tactical and strategic decisions.

Introduction:

The global retail empire Walmart has attracted millions of customers from all over the world by offering a broad selection of affordable products. In today's digital world, data-driven insights are becoming essential for companies like Walmart to remain competitive as consumer behavior changes. Through the examination of Walmart's extensive customer demographic data and their purchase history, this research seeks to identify patterns that may inform more specialized and successful marketing initiatives or strategies. Walmart can increase its operational efficiency and improve customer service of its diverse customer base by analyzing their customer data and behaviors.

Predictive modeling & Data analysis techniques will be applied to forecast purchasing behavior and estimate future trends based on historical data. Additionally, clustering techniques will help

identify distinct customer segments based on factors such as age, city category, and marital status. These methods will allow us to explore how variables such as occupation and years spent in a city impact spending behavior.
The insights gained from this analysis will help Walmart's business strategy to be more in line with customer needs.

Walmart now can tailor its product assortments, develop targeted marketing campaigns, and improve their customers engagement by identifying patterns of purchases and preferences of their customers. In the end, by staying ahead of consumer trends and expectations, this project will help Walmart maintain its competitive edge in the retail sector while also enhancing customer satisfaction and loyalty.

Dataset Description:

The dataset consists of detailed information about 550,068 Walmart customer transactions and their purchasing behavior, containing various demographic details and product-related data. Each row in the dataset represents an individual transaction, allowing for the aggregation of purchases across various demographic segments, such as gender, age, and marital status. These insights can aid in developing targeted marketing strategies, improving product recommendations, and optimizing inventory management.

Column Name	Type	Description
User_ID	Categorical	Unique identifier for each customer.
Product_ID	Categorical	Unique identifier for each product purchased by the customer.
Gender	Categorical	The gender of the customer (Male/Female).
Age	Categorical	The customer's age, categorized into predefined age bins.
Occupation	Numerical (Discrete)	A masked identifier representing the customer's occupation.
City_Category	Categorical	The category of the city where the customer resides (A, B, or C).

StayInCurrentCityYears	Categorical	The number of years the customer has lived in their current city.
Marital_Status	Binary	Indicates the marital status of the customer (Married/Not Married).
ProductCategory	Categorical	A masked identifier representing the product category.
Purchase	Numerical(continuous)	The amount of money the customer spent on a particular product.

Target Variable:

In the given dataset, the focus variable will be the "Purchase" column, which shows how much money each client spends on goods. This variable would lead to a number of strategic applications required in demand forecasting, tailored marketing, and price sensitivity analyses, hence studying the expenditure trends and forecast future purchases. In this regard, Walmart can enhance the decision-making process related to effective inventory management, customer segmentation, and sales optimization by focusing on the amount under the "Purchase" column. This will also give Walmart more comprehension regarding client preferences.

Methodology:

Data Collection and Loading:

- Import and load the Walmart sales dataset for customer demographics and purchase records.

Data Cleaning:

- Check the dataset for missing values, duplicates, and inconsistencies.
- Change the data type of applicable columns to 'category' to handle it correctly for the analysis.

Exploratory Data Analysis (EDA):

- Demographic features to be analyzed: gender, age, occupation, city category, and marital status.
- Visualize the distribution, pattern, and correlation using count plots, bar charts, and histograms.
- Identify the trend in purchase behaviour in various customer segments to find out those groups who make high purchases.

Customer Segmentation:

- K-Means Clustering**:** Use the K-Means Clustering algorithm to divide up your clientele into groups according to their demographics and purchase patterns. To develop focused marketing tactics, use cluster analysis to discover particular client segments (such as high spenders and frequent shoppers).
- Using hierarchical clustering, you may find nested groupings in the customer data and get a more in-depth understanding of the consumer segments according to their purchase habits.
- Density-Based Spatial Clustering of Applications with Noise, or DBSCAN, is a technique that can detect outlier consumer segments that might need particular approaches by identifying core customer segments based on the density of purchasing behaviour.

Predictive Modeling:

- Regression Analysis:
    - Use Multiple Linear Regression to predict sales based on independent variables such as customer demographics and purchase history.
    - Discuss various features affecting purchasing habits, then identify the significant predictors.
- Decision Trees:
    - Decision Tree models represent a classification of customers based on their demographic factors and likelihood to make a purchase, using a tree-like model of decisions and their possible consequences.
- Random Forest:
    - Apply Random Forest to improve the accuracy of predictions and interpret the importance of features for customer behaviour prediction tasks with several input variables.

Generate Insights:

- Customer Demographics: Determine, from demographic factors such as age, sex, and level of income, those that would correspond to higher purchasing rates for certain products. Describe how each target segment may respond differently to different marketing strategies.
- Seasonal Trends: Analyze temporal patterns in purchasing to find seasonality or events that have the greatest impact on the popularity of the products. This will help in determining the best timing for focused promotions and inventory management.

Tools And Technologies:

1. Data Collection & Preprocessing:
    - Python: Primary programming language for data handling.
    - Pandas: Data manipulation and analysis (cleaning, merging, filtering data).
    - NumPy: Numerical computing for handling arrays and performing calculations.
2. Exploratory Data Analysis (EDA):
    - Matplotlib: Basic plotting and data visualization.
    - Seaborn: Advanced data visualization (heatmaps, categorical plots).
    - Jupyter Notebook: Interactive environment for analysis and visualization.
3. Descriptive & Predictive Analytics:
    - Scikit-learn: Machine learning models (linear regression, random forests, decision trees, etc.), model evaluation, and cross-validation.
    - Statsmodels: For statistical analysis and regression modeling.
4. Visualization & Reporting:
    - Tableau or Power BI: For interactive dashboards and visual reporting.
    - Matplotlib/Seaborn: Static visualizations and graphs.
5. Collaboration & Version Control:
    - Git/GitHub: Version control and collaborative coding.
    - Google Colab: Cloud-based collaborative notebooks.

Research Questions:

1. How do customer demographics (age, gender, marital status) influence purchasing behaviour at Walmart?
    - Investigate the relationship between demographic characteristics and total spending or product preferences.
2. What are the key factors driving high-value purchases among Walmart customers?
    - Identify which factors (e.g., age, city category, occupation) are most strongly associated with larger purchase amounts.
3. How does customer retention vary based on years spent in the current city?
    - Analyze if the number of years a customer stays in their city affects their purchasing loyalty and frequency.
4. Can we predict future purchasing behaviour based on customer profiles and past transactions?
    - Explore if demographic, geographic, and past purchasing data can effectively predict a customer's future purchase categories or amounts.
5. How does Walmart's product demand vary across different city categories (A, B, C)?
    - Determine if geographic factors significantly impact product preferences and demand patterns.
6. Which product categories are most likely to see an increase in demand based on current trends?
    - Identify patterns in purchasing behaviour that can forecast rising demand for specific product categories.
7. Does marital status influence the types of products purchased or the total spending at Walmart?
    - Explore if married versus single customers have differing product preferences and spending patterns.
8. What role does occupation play in determining spending patterns at Walmart?
    - Investigate how customers' masked occupations relate to their purchase amounts and product category preferences.

Project Outcome:

The focus of the project is to predict Walmart sales data, considering the customer demographic information such as Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, and Marital_Status, along with product information like Product_ID and Product_Category. The target variable 'Purchase' would give the amount of purchase based on those factors. It can use insights about demographics and product types to make pretty effective data-driven decisions in the area of inventory and marketing strategies. Different machine learning models for more accurate predictions have been implemented, which include Linear Regression, Random Forest Regressor, and Decision Trees. Both linear and nonlinear relationships between independent variables and purchase amounts are modeled through this. The outcome of this project will not be confined to the prediction of purchase amount alone but also to some key aspects such as demographics that influence spending behavior among the customers. This will help Walmart in providing marketing campaigns, improved pricing techniques to increase sales transactions, and personalized user product recommendations. Analysis can also help in understanding occupational and geographical mix on purchase patterns for efficient business strategies.