Project Proposal: Predictive Analytics for Customer Behavior in E-commerce

Team Name: OG Data Wizards

Team Members

- Venkata Sai sandeep Nirujogi - S564888
- Harichaithanya Kotapati - S567067
- Teja kumar Muppala - S5565960
- Riyaz Hussain Shaik - S567058

Project Title

Predictive Analytics for Customer Behavior in E-commerce

Project Idea

The goal of this project is to analyze and predict customer behavior in the e-commerce domain using the Online Retail dataset. This dataset contains transactional data from an online retail store, including fields like product descriptions, quantities, prices, and customer information. By analyzing patterns in customer purchase history, we will develop predictive models to forecast customer behavior, such as future purchases, churn likelihood, product preferences, and seasonal sales trends.

Tools and Technologies

- **Programming Languages:** Python, SQL
- **Data Processing Frameworks:** Apache Spark
- **Big Data Storage and Processing:** Hadoop, Apache Kafka
- **Data Visualization:** Matplotlib, Seaborn, Plotly
- **Version Control:** Git/GitHub for collaborative development

- **Dataset:** Online Retail Dataset from UCI Machine Learning Repository
- **Reference Link:** Online Retail Dataset - UCI Repository

High-Level Architecture

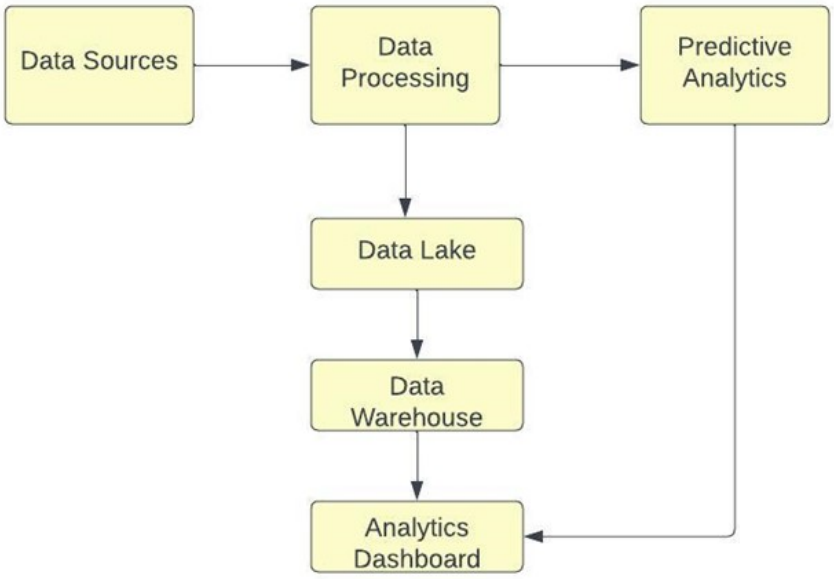

Figure 1: High-Level Architecture Diagram for Predictive Analytics System

Explanation of the Diagram:

- **Data Sources:** The data comes from various e-commerce systems, including transactional databases, CRM systems, and website logs. Additionally, external sources like customer feedback, social media interactions, and seasonal trends can be incorporated into the analysis.
- **Data Processing:** Raw data undergoes processing using Apache Spark for tasks like data cleaning, aggregation, and feature engineering. We perform operations such as handling missing values, aggregating sales data, and calculating metrics like customer lifetime value.
- **Data Warehouse:** Processed and transformed data is then organized into a Data Warehouse. This structured data is optimized for efficient query execution and reporting, making it ideal for running machine learning models and conducting deep-dive analyses.

- **Dashboard/Visualization Layer:** Data visualizations are built using Matplotlib and Seaborn to present insights in the form of charts, graphs, and interactive dashboards. The visualizations will be used to monitor trends, customer segmentation, and product performance.
- **Analytics Dashboard:** Predictive insights are visualized through an analytics dashboard, providing actionable insights to stakeholders.

Goals

1. Identify different customer segments based on purchase behavior, frequency, and product preferences.
2. Forecast future sales by predicting demand for specific products based on historical trends, seasonality.
3. Customer Segmentation by Total Spending aims to group customers based on their total spending behavior using K-Means Clustering. By aggregating total spending for each customer, we can identify high-value, moderate-value, and low-value customer segments.
4. Products with Highest Return Rates aims to identify products with the highest return frequency by analyzing the ratio of returned items to total sold items.
5. Monthly Sales Trends aims to analyze and visualize sales patterns over time, identifying seasonal fluctuations and trends.