Milestone 1: Global E-Commerce Review Analytics

Introduction

This project implements a comprehensive machine learning pipeline to predict the sentiment group of e-commerce product reviews based on user demographics, product characteristics, and review-related features. The application addresses a real-world business problem: understanding customer satisfaction patterns and product performance across regions to improve recommendation systems, marketing strategies, and logistics efficiency.
E-commerce platforms, when analyzed effectively, provide invaluable insights into consumer behavior, the impact of delivery performance, and the influence of product attributes on user satisfaction. By leveraging data analytics, online retailers can personalize recommendations, optimize delivery processes, and improve customer retention, ultimately enhancing revenue and loyalty.

Dataset

1. Products Dataset

The products.csv file provides product-specific attributes including `product_id, product_category_name, product_name_length, product_description_length, product_photos_qty, product_weight_g, product_length_cm, product_height_cm, and product_width_cm`. These attributes are essential for understanding how product characteristics (such as size, photos, or descriptions) influence review sentiment and perceived quality.

2. Reviews Dataset

The reviews.csv file serves as the central component of the dataset, representing individual customer feedback. Each row corresponds to one review containing both structured and unstructured information. Key columns include `review_id, order_id, review_score,`

`review_comment_title, review_comment_message, and review_creation_date`. The rating scale (1–5) captures customer satisfaction across multiple dimensions, enabling sentiment classification into Positive, Neutral, or Negative groups.

3. Order Items Dataset

The `order_items_dataset.csv` file serves as the bridge between orders and products. It contains `order_id`, `product_id`, `price`, `freight_value`, `seller_id`, `shipping_limit_date.` Each record represents one item within an order, allowing the connection between **products purchased** and **order details**.

4. Customers Dataset

The customers.csv file contains demographic and location-related information such as `customer_id, customer_city, and customer_state.` These details are indispensable for regional analysis and identifying market-specific behavior patterns.

5. Orders Dataset

The orders.csv file records transactional information including `order_id, customer_id, order_status, order_purchase_timestamp, order_delivered_customer_date, and order_estimated_delivery_date`. This dataset provides delivery-related insights—critical for understanding the effect of logistics delays on review sentiment.

Dataset Source: [Brazilian E-Commerce Public Dataset by Olist | Kaggle]

Project Objectives

1. Data Cleaning

Remove irrelevant or redundant columns, handle null or missing values, convert timestamp columns to datetime format, and compute additional derived attributes (e.g., delivery delay, review length). Ensure the 5 datasets are merged correctly through shared keys (`customer_id, order_id, and product_id`).

2. Data-Engineering Questions

From the combined dataset, analyze and visualize the reasoning for the following questions:

- Which product category has the highest average review score in each region?
- What is the relationship between delivery delay and review sentiment?
- Which regions show the highest proportion of negative reviews?

3. Predictive Modeling Task

Develop a statistical ML model or shallow feedforward neural network (FFNN) to predict the sentiment group of a review (Positive, Neutral, or Negative) based on the following features:

- Score-Based Features derived from the review rating and sentiment extracted from review text.
- Product Features including description length, photo quantity, and category.
- Customer Features such as region or state.
- Delivery Features capturing delivery delay and status.
- Price Features: item price and freight value.

This is a multi-class classification problem. Models will be evaluated using accuracy, precision, recall, and F1-score.

4. Model Explainability

To gain deeper insights into the model's behaviour, apply explainable AI (XAI) techniques—specifically SHAP or LIME—to interpret predictions, ensuring transparency

about the most influential features. Explainability will help uncover which factors (delivery time, product category, review length, etc.) drive customer sentiment, supporting data-driven recommendations for improving customer satisfaction.

Sentiment Groups

Sentiment Group	Review Scores	Description
Positive	4 – 5	High customer satisfaction
Neutral	3	Average or mixed feedback
Negative	1 – 2	Dissatisfied customers

Project Deliverables

1. A refined Jupyter Notebook documenting all steps with explanations and visualizations.
2. A cleaned dataset containing engineered features and a new sentiment_group column.
3. An analytical report answering the data-engineering questions with clear visual reasoning.
4. A predictive model (statistical ML or FFNN) to classify sentiment groups.
5. XAI outputs (SHAP plots or LIME explanations) illustrating the contribution of different features to predictions.
6. An inference function that takes new review and order data and returns the model's predicted sentiment group.
7. A business insights report summarizing key findings and actionable recommendations.

Submission and Deadline

Please submit your GitHub repository containing all project files, fulfilling the specified requirements, using the course submission form. Ensure your repository remains private until the deadline. Afterward, you may make it public or add the course account as a collaborator for grading.