Machine Learning: Activity - Data Preprocessing

Vikas Thammanna Gowda

01/30/2025

Problem Description: Local Grocery Store Sales Data

A small, family-owned grocery store has been operating in a residential neighborhood for the past 10 years. The store sells a variety of products, including fresh produce, packaged goods, dairy, and household essentials. The owner has been manually recording sales data but now wants to use data-driven insights to improve inventory management, predict sales trends, and optimize pricing strategies.

The store has provided transaction-level sales data recorded over the past two years. However, before analyzing the data or building predictive models, data preprocessing is necessary to clean and standardize the dataset.

Dataset Overview

The following table provides a description of the dataset collected by the store. **Provide the datatypes.**

Feature Name	**Description**	**Data Type**
Transaction ID	Unique identifier for each transaction	
Date & Time	Timestamp of when the transaction occurred	
Day of the Week	Day on which the transaction occurred (Monday-Sunday)	
Product Name	Name of the product purchased	
Product Category	Category the product belongs to (e.g., Fresh Produce, Dairy, Snacks)	
Product Price	Price of a single unit of the product	
Quantity Sold	Number of units of a product purchased in the transaction	
Total Transaction Amount	Total amount spent in the transaction	
Discount Applied	Discount given in the transaction (if any)	
Payment Method	How the customer paid (Cash, Credit Card, Mobile Payment)	
Customer Type	Whether the customer is a walk-in or a member of the store's loyalty program	
Stock Level at Purchase	The available stock of the product at the time of purchase	
Supplier Name	The supplier/vendor of the product	
Weather Condition	Weather at the time of purchase (Sunny, Rainy, Snowy, etc.)	
Neighborhood Event	Whether a local event (e.g., school fair, community gathering) was happening nearby	
Expiration Date	Expiration date of perishable products	
Restock Frequency	How often the store replenishes stock for this product	

Table 1: Dataset Features Description

Discussion Questions

Feature Selection

- Looking at the dataset, discuss which features might be most relevant for analysis and which ones may not add significant value.

Outlier Treatment

- From the selected features, which ones are more likely to have outliers?

- What are possible sources of outliers in the data? (e.g., incorrect data entry, bulk purchases, extreme discounts).

- How would you decide whether an outlier should be removed, corrected, or retained?

- What types of visualizations could help in identifying outliers?

Feature Scaling

- Are there features in the dataset that have different numerical ranges and may need to be scaled for consistency?

- If scaling is applied, should it be applied to all numerical features, or should some be left unchanged?

Feature Extraction / Engineering

- Are there features that could be combined or transformed to create a more useful representation?

- How might categorical features be processed to ensure they are useful for analysis and modeling?