Data Collection and Preprocessing Phase

Date	5 July 2024
Team ID	SWTID1720082658
Project Title	Ecommerce Shipping Prediction Using Machine Learning
Maximum Marks	2 Marks

Data Collection Plan & Raw Data Sources Identification Template

We've designed this template to help our team gather the essential data for predicting shipping times in e-commerce using machine learning. It provides a clear roadmap for identifying and organizing raw data sources, ensuring we capture key details like order info, shipping durations, customer data, and logistics. The focus is on collecting high-quality, relevant, and consistent data for an accurate prediction model. Key sections guide us through sourcing, collecting, validating data, and addressing any integration challenges.

Data Collection Plan Template

Section	Description
Project Overview	The E-Commerce Shipping Prediction project aims to forecast shipping times using machine learning. Key steps include collecting and preprocessing data on orders, customers, and shipping logistics, performing exploratory data analysis, selecting and training regression or classification models, and deploying the model for real-time or batch predictions. The project seeks to provide accurate delivery estimates, optimize shipping operations, and enhance customer satisfaction while addressing challenges like data quality and model scalability.
Data Collection Plan	The dataset was obtained from the E-Commerce Shipping Data from Kaggle.

Raw Data Sources Identified	• ID: ID Number of Customers. • Warehouse block: The Company have big Warehouse which is divided in to block such as A,B,C,D,E. • Mode of shipment:The Company Ships the products in multiple way such as Ship, Flight and Road. • Customer care calls: The number of calls made from enquiry for enquiry of the shipment. • Customer rating: The company has rated from every customer. 1 is the lowest (Worst), 5 is the highest (Best). • Cost of the product: Cost of the Product in US Dollars. • Prior purchases: The Number of Prior Purchase. • Product importance: The company has categorized the product in the various parameter such as low, medium, high. • Gender: Male and Female. • Discount offered: Discount offered on that specific product. • Weight in gms: It is the weight in grams. • Reached on time: It is the target variable, where 1 Indicates that the product has NOT reached on time and 0 indicates it has reached on time.

Raw Data Sources Template

Source Name	Description	Location/URL	Format	Size	Access Permissions
E-Commerce Shipping Data - Kaggle	The dataset used for model building contained 10999 observations of 12 variables.	https://www.kaggle.com/datasets/prachi13/customer-analytics?select=Train.csv	CSV	430 KB	Public