Dmart Customer Segmentation with Machine Learning

This presentation outlines a machine learning project focused on customer segmentation using Dmart sales data. The aim is to understand customer behavior by grouping them based on purchasing patterns, product categories, quantities, and spending habits. Analyzing these perspectives provides valuable insights for targeted marketing and business strategy optimization. We'll explore the workflow, data preprocessing, cluster analysis, and model training steps involved in this project.

by Sankari C

Project Workflow Overview

1

Dataset Acquisition

The project begins with acquiring the Dmart sales dataset, which contains transactional information about customer purchases, product details, and quantities.

2

Data Preprocessing

Raw data is cleaned, transformed, and prepared for analysis. This involves handling missing values, feature selection, and data type conversions.

3

Data Analysis

Exploratory data analysis (EDA) is conducted to understand the distribution of features, identify patterns, and gain insights into customer behavior.

4

K-Means Clustering

The K-means algorithm is trained on the preprocessed data to group customers into distinct clusters based on their purchasing behavior.

5

Visualization and Analysis

The resulting clusters are visualized and analyzed to identify patterns, characteristics, and relationships between customer segments and product categories.

Loading and Inspecting the Dataset

Loading the Data

The initial step involves loading the Dmart sales dataset into a suitable data structure, such as a Pandas DataFrame, for further processing.

First 5 Rows

Printing the first 5 rows of the DataFrame to get a glimpse of the data structure and the nature of the data.

Checking the Shape

Determining the number of rows and columns in the dataset to understand its size and dimensionality.

Data Preprocessing Steps

Handling Missing Values

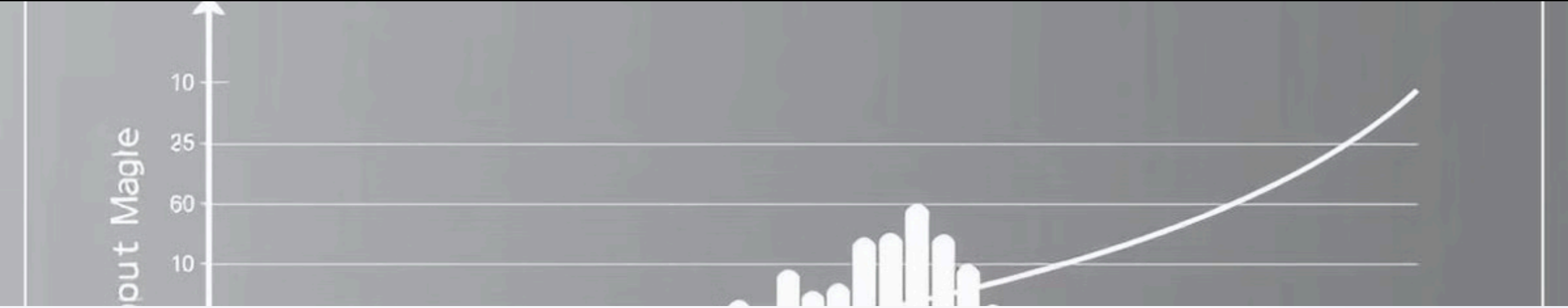
Identifying and addressing missing values in the dataset, either by imputation or removal, to ensure data quality and integrity.

Feature Selection

Selecting relevant features for clustering, such as product categories, unit prices, quantities, and total prices, while excluding irrelevant or redundant features.

Data Slicing

Slicing the DataFrame to extract the selected features into a new data structure for subsequent standardization and clustering.



Standardizing the Data

Standardizing the selected features using techniques like StandardScaler to ensure that all features have a similar scale and distribution. This prevents features with larger values from dominating the clustering process.

Standardization involves transforming the data so that it has a mean of zero and a standard deviation of one. This ensures that each feature contributes equally to the distance calculations in the clustering algorithm.

By standardizing the data, we improve the accuracy and reliability of the clustering results, leading to more meaningful customer segments.

Finding the Optimal Number of Clusters

WCSS Calculation

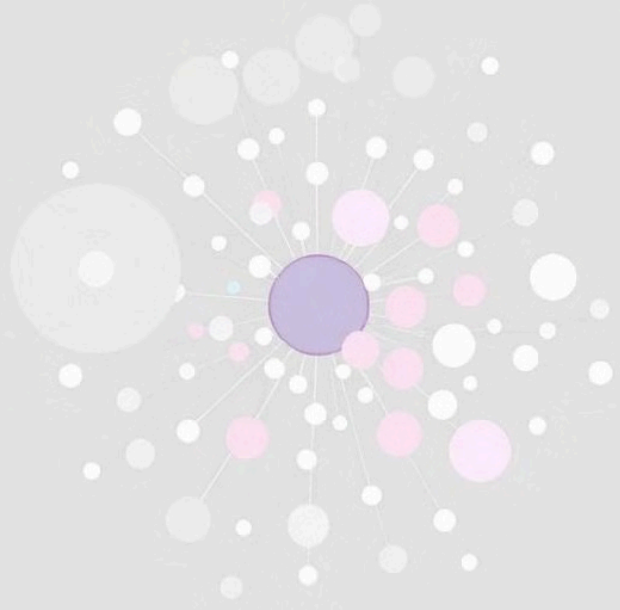
WCSS (Within-Cluster Sum of Squares) measures the sum of squared distances between each data point and its assigned cluster centroid. Lower WCSS values indicate tighter clusters.

Elbow Method

The elbow method involves plotting WCSS values against different numbers of clusters (k) and identifying the "elbow" point where the rate of decrease in WCSS starts to diminish. This point represents the optimal number of clusters.

Visual Inspection

Visualizing the elbow graph allows for easy identification of the optimal cluster point, providing a visual representation of the trade-off between the number of clusters and the tightness of the clustering.



Training the K-Means Clustering Model

1

Initialization

Initialize the K-means clustering algorithm with the determined optimal number of clusters (k).

2

Model Fitting

Fit the K-means model to the standardized data, allowing the algorithm to learn the cluster centroids and assign data points to the nearest cluster.

3

Cluster Assignment

Assign each data point (customer) to its corresponding cluster based on the learned cluster centroids.

Analyzing and Visualizing Clusters

Cluster Profiling

Analyze the characteristics of each cluster, such as average spending, product preferences, and purchase frequency, to understand the behavior of different customer segments.

Visualization

Visualize the clusters using scatter plots visualization techniques to gain insights into the relationships between features and cluster assignments.

Relationship Analysis

Explore the relationships between cluster assignments and feature values to identify patterns and trends that differentiate customer segments by pairplots.