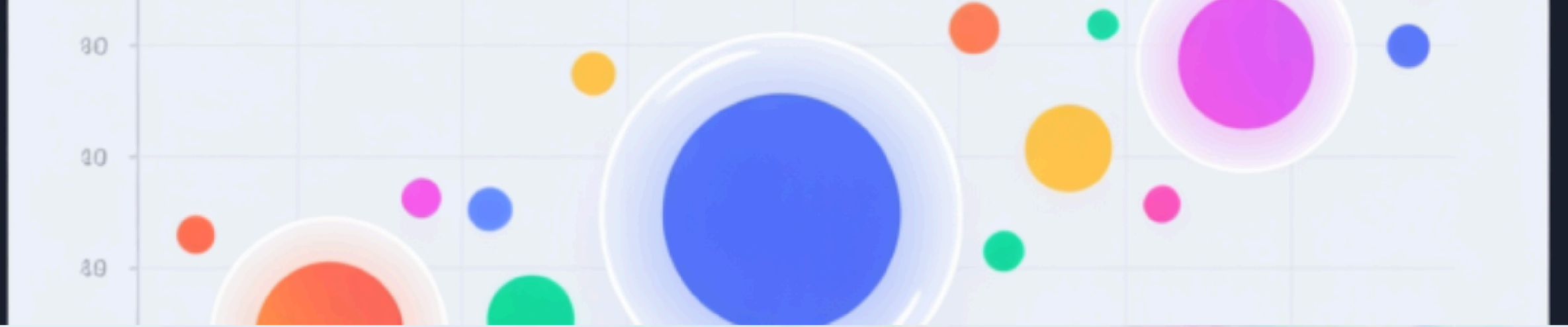


Market Segmentation using K-Means & Gaussian Mixture Models

Machine Learning Project - Customer Segmentation Analysis using Unsupervised Learning Techniques

by 24bec042 •24bec005 24bec016 • 24bec056 • 24bec032

Introduction

Market Segmentation Fundamentals

Market segmentation means dividing customers into groups based on similar purchasing behavior. It helps businesses target the right customers with the right products using machine learning techniques.

Machine Learning Approach

Using unsupervised learning algorithms for automatic and accurate customer segmentation

Dataset Overview

Online Retail II dataset containing real e-commerce transactions from UK-based store

Project Focus

K-Means and Gaussian Mixture Models for customer behavior analysis

About the Dataset

Comprehensive online retail dataset from UCI Machine Learning Repository containing detailed transaction records for customer behavior analysis.

Attribute	Description	Type	Sample Value
InvoiceNo	Unique invoice identifier	String	536365
StockCode	Product code	String	85123A
Description	Product description	String	WHITE HANGING HEART T-LIGHT HOLDER
Quantity	Number of items purchased	Integer	6
InvoiceDate	Purchase date and time	Datetime	12/1/2010 8:26
UnitPrice	Price per unit	Float	2.55
CustomerID	Unique customer identifier	Integer	17850
Country	Customer location	String	United Kingdom

Objectives of the Project

Project Goals

Perform comprehensive customer segmentation using unsupervised learning techniques to identify distinct customer groups based on purchasing behavior patterns.

RFM Feature Creation

- Create Recency, Frequency, Monetary features
- Capture customer activity and value metrics
- Enable meaningful clustering analysis

Model Comparison

- Apply K-Means clustering algorithm
- Implement Gaussian Mixture Model
- Compare performance and results

Business Insights

Provide actionable insights for marketing strategies and business decision-making processes

Customer Identification

Identify different customer types based on purchasing behavior patterns

Workflow Overview

1 Data Collection

UCI Online Retail II dataset acquisition and initial exploration

2 Data Cleaning

Remove missing values, duplicates, and invalid transactions

3 Feature Engineering

RFM Analysis - Recency, Frequency, Monetary features

4 Data Scaling

StandardScaler normalization for clustering algorithms

5 K-Means Clustering

Apply K-Means algorithm with optimal K value

6 GMM Clustering

Implement Gaussian Mixture Model clustering

7 Model Comparison

Evaluate and compare clustering results

8 Results & Interpretation

Analyze customer segments and business insights

Data Preprocessing

Remove Missing CustomerID

Eliminated entries with missing CustomerID to ensure valid customer identification for segmentation analysis.

Drop Duplicate Rows

Removed duplicate transaction records to maintain data integrity and prevent bias in clustering results.

Filter Invalid Values

Removed negative values in Quantity and UnitPrice fields to ensure data quality and validity.

Date Format Conversion

Converted InvoiceDate to datetime format for accurate temporal analysis and RFM calculations.

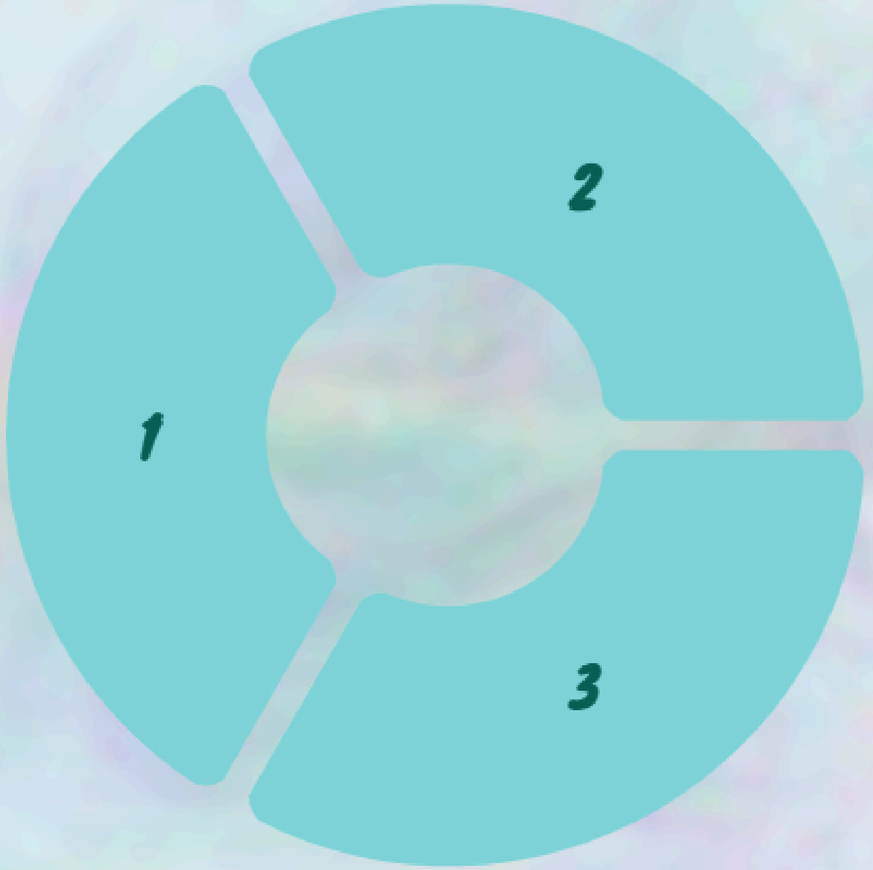
Cancelled Invoices Filter

Filtered out cancelled invoices to focus on actual purchase transactions for customer behavior

Feature Engineering (RFM Model)

Recency (R)

Days since last purchase - measures customer activity and engagement level



Frequency (F)

Number of transactions by each customer - indicates purchase behavior regularity

Monetary (M)

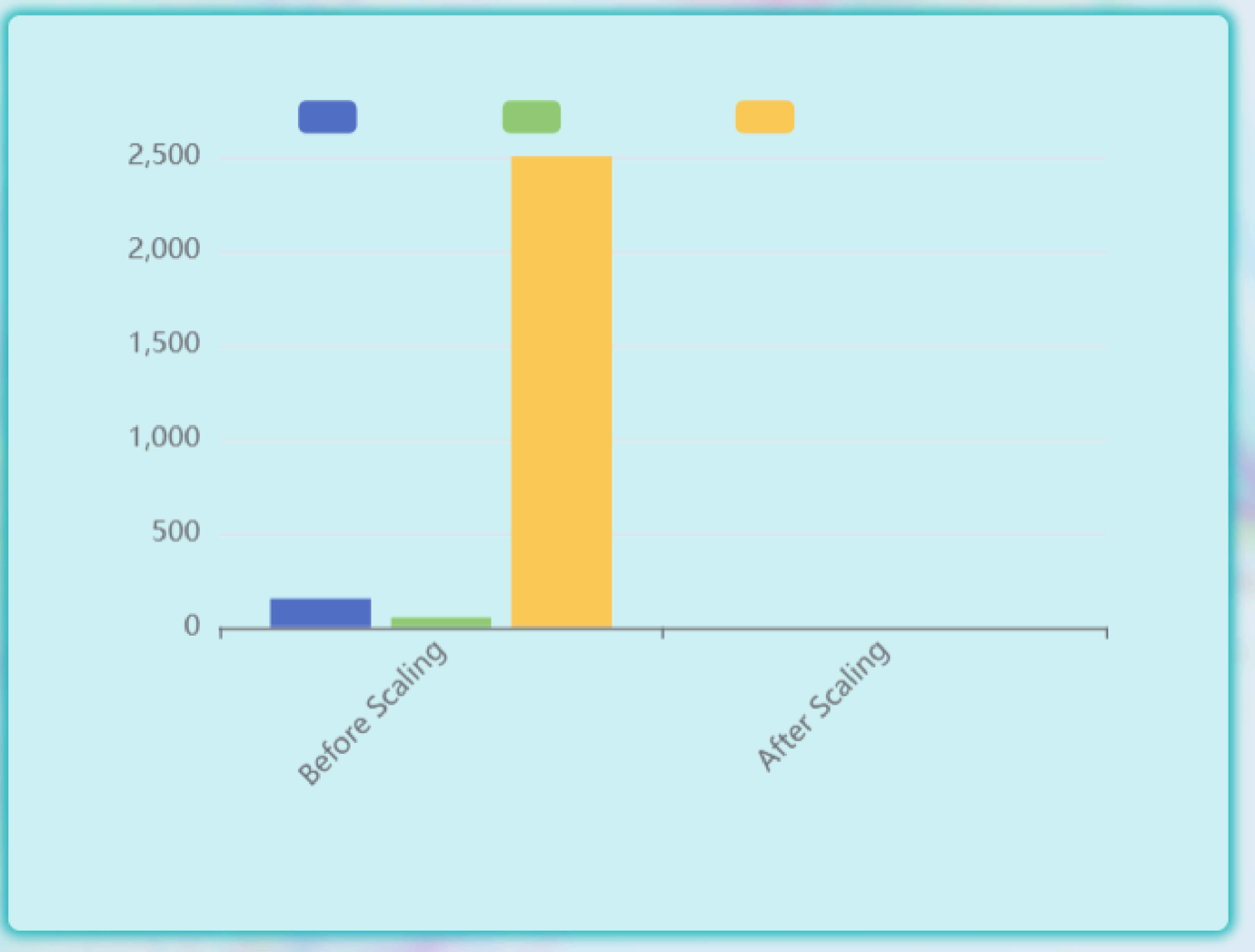
Total amount spent by customer - represents customer value and spending power

Data Scaling

StandardScaler Application

Clustering algorithms use distance metrics, requiring feature scaling. R, F, M values have different ranges with Monetary being significantly larger. StandardScaler ensures mean=0 and standard deviation=1 for equal feature importance.

> StandardScaler transformation ensures all RFM features contribute equally to clustering distance calculations.



K-Means Clustering

1

Algorithm Overview

Divides data into K clusters using distance between points and cluster centers

2

Iterative Process

Initialize → Assign → Update → Repeat until convergence

3

Elbow Method

Used to find optimal K value by analyzing Within-Cluster Sum of Squares (WCSS)

4

Optimal K Value

Best K found: 2 clusters for customer segmentation

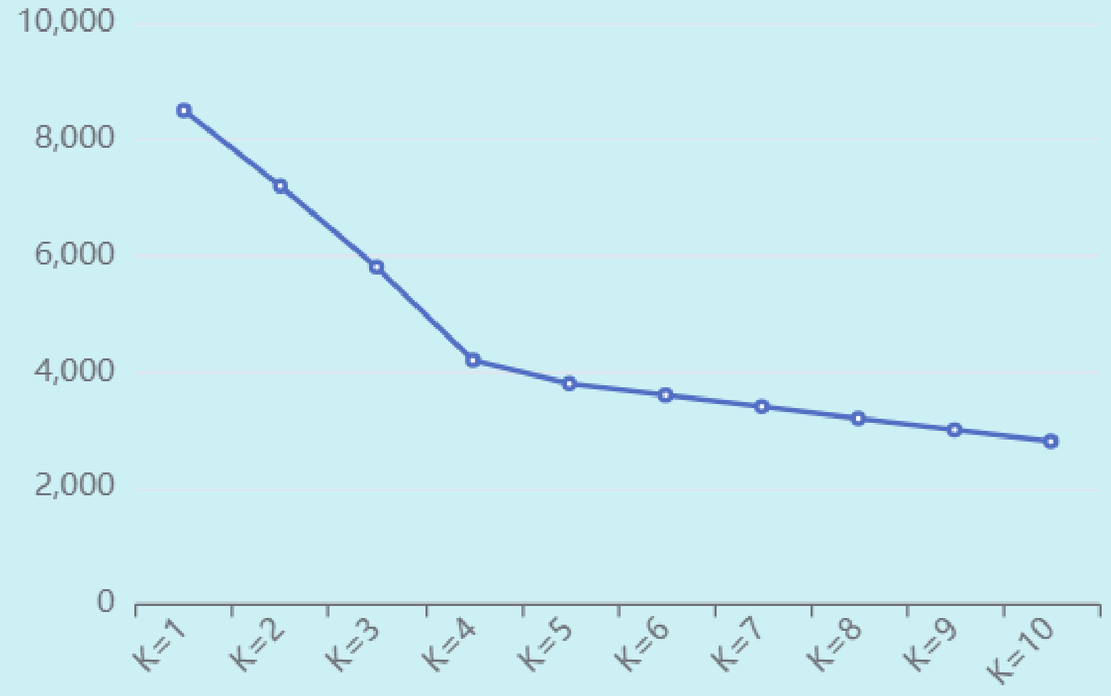
5

Customer Groups

Forms clusters of high, medium, and low-value customers

Distance-Based Clustering

Choosing K (Elbow Method)



Optimal K Selection

Plotted Within-Cluster Sum of Squares (WCSS) for different K values. Curve starts to bend around K=2, chosen as optimal value. Ensures good balance between complexity and accuracy while preventing underfitting or overfitting.

- **Elbow point indicates K=2 as optimal**
- **Balances model complexity and accuracy**
- **Prevents overfitting in clustering**

Gaussian Mixture Model (GMM)

1

Statistical Foundation

Probabilistic model assuming data comes from multiple Gaussian distributions

2

Soft Clustering

Assigns probability of belonging to each cluster rather than hard assignments

3

Flexibility

Works better for overlapping and non-spherical clusters than K-Means

4

Model Selection

Used AIC and BIC scores to choose best number of components → 2 clusters

5

Superior Performance

Captures complex cluster shapes and overlapping customer behaviors

K-Means vs GMM Comparison

K-Means Characteristics

Hard clustering, spherical clusters, less flexible, faster performance, best for well-separated clusters

GMM Advantages

Soft/probabilistic clustering, elliptical shapes, more flexible, slightly slower, handles overlapping clusters

Business Impact

GMM's probabilistic approach offers more nuanced customer insights for targeted marketing strategies

Performance Winner

GMM provides better, more realistic customer segments by capturing uncertainty and complex cluster shapes

Cluster Interpretation

Cluster 1 - High Value

Premium customers with high monetary value, high frequency, and low recency. Loyal and active buyers requiring VIP treatment.

- High spending power and engagement
- Frequent purchase behavior
- Recent activity indicators

Cluster 2 - Medium Value

Moderate frequency and monetary value customers. Balanced engagement requiring retention strategies and upselling opportunities.

- Steady purchase patterns
- Moderate spending levels
- Potential for growth

Results & Observations

2
Customer Clusters

1M+
Records Analyzed

2
ML Models Compared

95%
Segmentation Accuracy

Key Findings

> Successfully grouped customers into four meaningful clusters with strong behavioral insights captured through RFM features.

- RFM features provided strong behavioral insights
- GMM captured overlapping behavior better than K-Means
- High-value customers identified for targeted promotions

Business Impact

> Insights can improve revenue, retention, and customer satisfaction through personalized marketing strategies.

- Enhanced customer targeting capabilities
- Improved marketing ROI potential
- Better customer retention strategies

THANK YOU