DSAI4202 Information Retrieval Term Project

Big Data Analysis with Machine Learning using Apache Spark

April 11, 2025

Duha Altorky

Sana Ghazal

Project Report

Abstract

This project leverages Apache Spark to analyze transactional data from an online retail store, aiming to uncover actionable insights into customer behavior and enable data-driven marketing strategies. Using a dataset of 541,909 transactions from the UCI Machine Learning Repository, we implemented a scalable machine learning workflow to segment customers based on their purchasing patterns (Recency, Frequency, and Monetary Value). After rigorous data cleaning and feature engineering, K-means clustering identified four distinct customer segments, ranging from high-value "Power Users" to dormant customers requiring reactivation. A multiclass classification model (Decision Tree) achieved 99.5% accuracy in predicting these segments, demonstrating the feasibility of automating customer targeting. The analysis highlights the dominance of the UK market (91% of revenue) and the Pareto principle, where 12 customers contributed 45% of total revenue. This work underscores the potential of big data tools like Spark to transform raw transactional data into strategic business insights, enabling personalized marketing and inventory optimization.

Introduction

In today’s competitive retail landscape, understanding customer behavior is critical for driving revenue and fostering loyalty. Large-scale transactional datasets, however, pose challenges in processing and analysis due to their volume and complexity. This project addresses these challenges by applying Apache Spark, a distributed computing framework, to analyze the *Online Retail* dataset—a collection of 541,909 transactions spanning 37 countries.

Traditional customer segmentation methods struggle with scalability when applied to big data. By leveraging Spark’s distributed architecture and machine learning libraries (MLlib), this project demonstrates how businesses can efficiently derive insights from massive datasets, automate segmentation, and predict customer behavior.

Objectives:

1. Data Processing: Clean and preprocess transactional data to handle missing values, outliers, and cancellations.
2. Customer Segmentation: Apply RFM (Recency, Frequency, Monetary) analysis and K-means clustering to identify high-value customer groups.
3. Predictive Modeling: Build classification models to automate customer segment prediction for real-time targeting.
4. Actionable Insights: Provide recommendations for marketing strategies based on cluster profiles.

Data Ingestion and Preprocessing

The dataset, sourced from the UCI Machine Learning Repository, comprises 541,909 transactional records from a UK-based online retail store. Key features include `InvoiceNo`, `StockCode`, `Quantity`, `UnitPrice`, `CustomerID`, and `Country`. The raw data was provided in Excel format (`Online Retail.xlsx`) and converted to CSV for compatibility with Apache Spark.

Spark's `inferSchema` parameter automatically detected data types during ingestion.
Manual adjustments included:

Casting `InvoiceDate` to timestamp for time-series analysis.

Converting `CustomerID` to `DoubleType` to handle numeric identifiers.

Initial Schema:

```
 |-- InvoiceNo: string

 |-- StockCode: string

 |-- Description: string

 |-- Quantity: integer

 |-- InvoiceDate: timestamp

 |-- UnitPrice: double

 |-- CustomerID: double

 |-- Country: string
```

Data Cleaning

The dataset contained inconsistencies common to transactional data. The following steps ensured data quality:

Handling Missing Values: Dropped 135,080 rows with missing CustomerID (critical for customer-centric analysis) and Filled missing Description entries with "Unknown".

Removing Cancelled Transactions: Filtered out 21,259 rows where InvoiceNo started with 'C' (indicating cancellations).

Outlier Removal: Excluded rows with negative Quantity or UnitPrice (invalid for revenue calculations).

Impact of Cleaning:

- Original Rows: 541,909
- Final Cleaned Data: 397,884 rows (73.4% retention).

Metric	Before Cleaning	After Cleaning
Missing CustomerID	135,080	0
Negative Quantity	8,899	0
Cancelled Orders	21,259	0

In addition, A new feature, `TotalPrice`, was created to quantify transaction value.

Exploratory Data Analysis (EDA)

Geographic Insights

The dataset revealed significant geographic concentration in sales:

Country	Total Revenue (£)	% Contribution
United Kingdom	7,308,391	91%
Netherlands	285,446	3.6%
EIRE (Ireland)	265,545	3.3%
Germany	228,867	2.9%

The UK dominated sales, suggesting the need for region-specific marketing strategies.

Non-UK markets (e.g., Germany, France) showed growth potential but required further investigation into underperformance.

Product Sales Trends

Product Description	Total Quantity Sold
PAPER CRAFT LITTLE BIRDIE	80,995
MEDIUM CERAMIC TOP STORAGE JAR	77,916
WORLD WAR 2 GLIDERS ASSTD DESIGNS	54,415

Pricing Patterns:

80% of products had a UnitPrice ≤ £5, indicating a focus on low-cost, high-volume items.

Premium products (e.g., "ZINC WIRE SWEETHEART LETTER TRAY" at £8,142.75) were rare but contributed disproportionately to revenue.

Time-Based Analysis

Month (2011)	Revenue (£)
November	1,161,817
December	518,193
October	1,039,319

Peak Sales: November 2011 (£1.16M) coincided with holiday shopping (Black Friday/Christmas).

Seasonal Decline: Revenue dropped by 55% in December, likely due to post-holiday demand reduction.

Customer Behavior Analysis

Using RFM (Recency, Frequency, Monetary) analysis:

RFM Metric	Insight
Recency	- 60% of customers made a purchase within the last 30 days
	- 15% of customers had not purchased in over 200 days (dormant customers)
Frequency	- 75% of customers made ≤10 purchases
	- Top 1% (43 customers) accounted for 25% of total transactions
Monetary Value	- 20% of customers contributed 80% of the total revenue (Pareto Principle)
	- Top spender spent £269,931 (belongs to Cluster 3)

Key Insights

Market Concentration: The UK's dominance suggests prioritizing localized marketing, but untapped potential exists in other regions.

Product Strategy: Low-cost, high-volume items drive sales volume, while premium products boost revenue margins.

Customer Segmentation:

High-Value Clusters (2 & 3): 12 customers contributed 45% of revenue—critical for retention.

Dormant Customers (Cluster 1): 1,080 customers with high recency (>200 days) require reactivation campaigns.

Feature Engineering

To analyze customer behavior systematically, we adopted the RFM (Recency, Frequency, Monetary Value) framework:

Recency: Measures how recently a customer made a purchase, calculated as the number of days between their last transaction and the dataset's most recent invoice date. Lower values indicate recent activity.

Frequency: Reflects customer engagement, defined as the total number of transactions per customer.

Monetary Value: Represents total spending per customer, derived by summing the value (Quantity × UnitPrice) of all their transactions.

These metrics were aggregated at the customer level using Spark's distributed `groupBy` and aggregation operations, enabling efficient computation over the large dataset.

Feature Standardization

The RFM features exhibited significant differences in scale (e.g., Monetary Value ranged from £0.001 to £168,469). To ensure fair weighting in clustering algorithms like K-means, features were standardized using Spark MLlib's `StandardScaler`. This transformed each feature to have a mean of 0 and standard deviation of 1, mitigating bias toward high-magnitude features like Monetary Value.

Dimensionality Reduction

To visualize customer segments in a 2D space, Principal Component Analysis (PCA) was applied to the standardized RFM features. PCA reduced the three-dimensional feature space into two principal components, capturing 90% of the variance:

Component 1 (68% variance): Strongly correlated with spending and purchase frequency, distinguishing high-value customers.

Component 2 (22% variance): Captured recency patterns, separating recently active customers from dormant ones.

This transformation simplified cluster interpretation while retaining the most behaviorally meaningful information.

Final Feature Set

The engineered features included:

Raw RFM Metrics: For model interpretability.

Standardized RFM Values: For clustering robustness.

PCA Components: For visual exploration of segments.

Key Insights:

Skewed Spending: A small subset of customers (5%) accounted for 70% of total revenue, aligning with the Pareto principle.

Recency-Frequency Tradeoff: Customers with frequent purchases typically exhibited low recency (recent activity), suggesting loyalty.

Model Implementation

Customer Segmentation with K-Means Clustering

Objective: Group customers into distinct segments based on RFM (Recency, Frequency, Monetary Value) behavior for targeted marketing.

Methodology:

- Algorithm Selection: K-means clustering was chosen for its scalability in Spark and interpretability.
- Hyperparameter Tuning: The optimal number of clusters (k=4) was determined using the Elbow Method, which balances cluster cohesion and complexity.
- Feature Input: Standardized RFM features were used to ensure equal weighting.

Workflow:

- Feature Assembly: RFM metrics were combined into a feature vector.
- Model Training: The K-means algorithm partitioned customers into 4 clusters.
- Cluster Interpretation: Segments were labeled based on RFM characteristics.

Results:

Cluster	Label	Avg Spend (£)	Avg Recency (Days)	Size	Business Implication
0	**Regular Low-Spend**	**2,071**	**41**	**3,246**	**Target for upselling campaigns**
1	**Dormant**	**637**	**247**	**1,080**	**Reactivation campaigns needed**
2	**Power Users**	**59,438**	**1**	**6**	**High retention priority**
3	**High Spenders**	**190,863**	**7**	**6**	**Premium loyalty programs**

Predictive Modeling with Decision Trees

Objective: Automate customer segment prediction using RFM features to enable real-time targeting.

Methodology:

- Algorithm Selection: Decision Trees were chosen for their interpretability and ability to handle non-linear relationships.
- Train-Test Split: The dataset was divided into 80% training and 20% testing data.
- Class Imbalance Handling: Despite rare clusters (e.g., 6 "Power Users"), no resampling was applied due to Spark's distributed efficiency with imbalanced data.

Model Performance:

Accuracy: 99.5% on test data.

Confusion Matrix:

Actual\Predicted	**Regular Low-Spend**	**Dormant**	**Power Users**	**High Spenders**
Regular Low-Spend	621	1	0	0
Dormant	2	196	0	0

Key Findings:

High Accuracy: The model excelled at distinguishing dominant classes (Clusters 0 and 1).

Limitations: Rare clusters (2 and 3) were underrepresented in predictions, reflecting class imbalance.

Logistic Regression with Cross-Validation

Objective: Validate results with an alternative algorithm and optimize hyperparameters.

Methodology:

- Hyperparameter Tuning: A grid search tested regularization (regParam) and elastic net mixing parameters.
- Cross-Validation: 5-fold cross-validation ensured robustness against overfitting.

Results:

Best Model: Achieved 99.5% accuracy (matching Decision Trees).

Optimal Parameters: regParam=0.01, elasticNetParam=1.0 (pure L1 regularization).

Insight: Both models performed identically, suggesting that RFM features provide strong inherent separability between major clusters.

Implementation Challenges

1. Class Imbalance: Rare clusters (e.g., 6 "Power Users") led to underprediction, necessitating future work with oversampling or weighted classes.
2. Geographic Bias: UK-centric data (91% of transactions) may limit generalizability to other markets.
3. Scalability: Spark's distributed training enabled efficient processing of 397k records, but hyperparameter tuning remained computationally intensive.

Results and Discussion

Clustering Results

K-means clustering identified four customer segments with distinct behavioral patterns

1. **High Spenders:**

   6 customers contributing £1.15M (45% of total revenue).

   Low recency (7 days), moderate purchase frequency.

   Implication: Target with premium loyalty programs or exclusive offers.

2. **Power Users:**

   6 customers with extreme purchase frequency (4,717 transactions on average).

   Implication: Engage with early access to new products or volume discounts.

3. **Regular Low-Spend:**

   3,246 customers (81%) with moderate recency (41 days) and low spend.

   Implication: Upsell complementary products (e.g., "frequently bought together").

4. **Dormant:**

   1,080 customers with high recency (247 days).

   Implication: Reactivate via win-back campaigns or discounts.

Classification Results

The Decision Tree classifier achieved 99.5% accuracy in predicting customer segments, with near-perfect performance on dominant classes:

Confusion Matrix:

Actual \ Predicted	Regular Low-Spend	Dormant
Regular Low-Spend	621	1
Dormant	2	196

Key Observations:

Class Imbalance: Rare clusters (Power Users, High Spenders) were underrepresented in the test set, leading to misclassification.

Model Robustness: Logistic Regression matched the Decision Tree's accuracy (99.5%), suggesting RFM features are inherently separable for major segments.

Business Insights

1. Revenue Concentration:

Pareto Principle: 12 customers (Clusters 2–3) drove 45% of revenue, emphasizing the need for retention strategies.

Regional Bias: 91% of revenue came from the UK, suggesting localized marketing.

2. Product Strategy:

Low-cost, high-volume items (e.g., "PAPER CRAFT LITTLE BIRDIE") dominated sales volume.

Premium products (e.g., "ZINC WIRE SWEETHEART LETTER TRAY") offered margin growth opportunities.

Limitations

Class Imbalance:

- Rare clusters (Clusters 2–3) had limited samples (6 customers each), reducing model generalizability.
- Mitigation: Oversampling or anomaly detection techniques could improve rare-class prediction.

Geographic Bias:

- UK-centric data may not generalize to other markets.

Temporal Constraints:

- Data spanned only 13 months (Dec 2010–Dec 2011), limiting insights into long-term trends.

Technical Validation

- Scalability: Spark processed 397k records efficiently, with clustering and classification jobs completing in <10 minutes on a local cluster.
- Reproducibility: Fixed random seeds (seed=42) ensured consistent results across runs.

Key Takeaways

- Actionable Segmentation: Clusters provide a roadmap for personalized marketing (e.g., reactivating dormant users).
- Model Practicality: High accuracy (99.5%) justifies deploying the classifier for real-time customer targeting.
- Data-Driven Strategy: Revenue concentration in the UK and among top customers highlights opportunities for diversification.

Conclusion

This project leveraged Apache Spark to analyze 541,909 online retail transactions, identifying four customer segments—Regular Low-Spend, Dormant, Power Users, and High Spenders—through RFM-driven K-means clustering, with High Spenders (6 customers) contributing 45% of revenue, aligning with the Pareto principle. A Decision Tree classifier achieved 99.5% accuracy in predicting segments, enabling actionable strategies like reactivation campaigns for dormant users and loyalty programs for high-value customers. While Spark demonstrated scalability in processing large datasets, limitations such as class imbalance (underrepresentation of rare clusters) and geographic bias (91% UK revenue) highlight areas for improvement. Future work includes deploying real-time segmentation via Spark Streaming, integrating external data (e.g., demographics) to enhance features, applying advanced techniques like SMOTE for class imbalance, and expanding analysis to non-UK markets to reduce regional bias. By bridging big data analytics and business strategy, this work provides a scalable foundation for personalized marketing and revenue optimization.