E-commerce Product Analysis

By
Arun K
Cohort ‘C’ – DADS

Project Objectives

Data collection and storage

The project aims to establish an end-to-end data pipeline for efficient data management and analysis.

Product Segmentation

Using unsupervised clustering techniques, to identify distinct Prosuct segments within the washing machine industry.

Business Insights

The ultimate goal is to derive actionable business insights that can inform strategic decision-making.

Problem Description

Unstructured Data

The large volume of unstructured e-commerce data makes analysis challenging and limits clear understanding of customer behavior.

Missing Values

Numerous records contain missing values, which can significantly impact the performance of machine learning models and analysis accuracy.

Understanding Market Segmentation

Washing machines differ by price, capacity, features, automation level, and build.

→ Clustering is needed to identify meaningful customer segments.

Predicting Product Segments

To train supervised ML models to classify washing machines into their segment categories based on their engineered features.

Workflow Overview

Web Scraping

BeautifulSoup and Selenium were utilized within Python to efficiently extract data from web sources.

Data Cleaning

Essential cleaning tasks included regex extraction, handling missing values, and feature engineering for better accuracy.

SQL Integraton

Storing cleaned data in a SQL database ensured structured access and efficient retrieval for analysis.

Web Scraping

- Web scraping was implemented to collect structured e-commerce data from multiple web sources.
- Python tools such as BeautifulSoup were used to extract key product details, including names, prices, ratings, and highlights, while handling pagination.
- This process ensured accurate and comprehensive data collection to support effective data analysis.

Data Cleaning

- Regular expression (regex) techniques were used to extract key attributes such as RPM, warranty information, and brand names.
- Additional product features were derived from product titles to enhance data completeness.
- This approach improved dataset quality and ensured more reliable analysis.

SQL Database Integration

- Cleaned data was systematically stored using SQLAlchemy to ensure reliable data persistence.
- The data upload process enabled seamless integration with the local database environment.
- This database setup supports efficient data retrieval and provides a strong foundation for advanced analysis and insights.

EDA PROCESS

price distribution

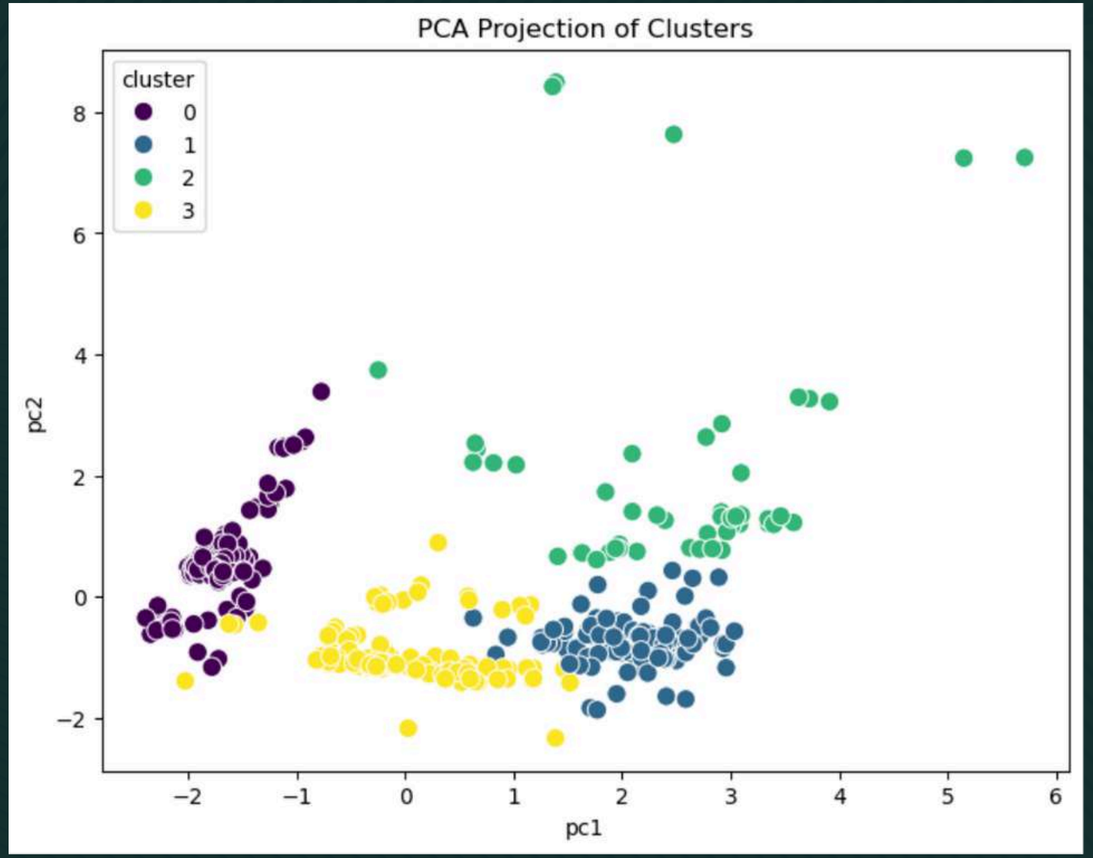
- It highlights the differences between budget and premium models, Understanding these variations is crucial for identifying pricing strategies

Correlation Analysis

- This analysis examines important relationships between price, capacity, heater influence, and RPM metrics.

Clustering Approach

- K-Means clustering was applied to group products into four clusters based on selected features.
- The elbow method was used to identify the optimal number of clusters for segmentation.
- The analysis revealed distinct market segments, highlighting differences in product features, price ranges, and customer preferences.



Cluster Summary

Average Price

The average price of each cluster reveals significant differences, indicating market diversity and consumer preferences.

Average Rating

Each cluster's average rating provides insights into customer satisfaction, guiding potential marketing strategies.

Demand

Demand metrics highlight trends in consumer behavior, pinpointing opportunities for targeted promotions and product offerings.

Segment Interpretation

- The washing machine market is segmented into Budget Semi-Auto, Mid-Range Top Load, Premium Front Load, and Flagship Front Load.
- Each segment differs in features, pricing, and customer preferences.

Supervised Models

Logistic Regression

```
---- Logistic Regression ----
Accuracy: 0.9457364341085271
F1 Macro: 0.9292984628833685
F1 Weighted: 0.946694202508156
```

kNN

```
---- kNN ----
Accuracy: 0.6666666666666666
F1 Macro: 0.5101226309921962
F1 Weighted: 0.6507306871310916
```

Random Forest

```
---- Random Forest ----
Accuracy: 1.0
F1 Macro: 1.0
F1 Weighted: 1.0
```

SVM

```
---- SVM ----
Accuracy: 0.6976744186046512
F1 Macro: 0.5356803327391563
F1 Weighted: 0.654299492876784
```

Decision Tree

```
---- Decision Tree ----
Accuracy: 0.9922480620155039
F1 Macro: 0.9876190476190476
F1 Weighted: 0.992366186784791
```

XGBoost

```
---- XGBoost ----
Accuracy: 0.9922480620155039
F1 Macro: 0.9859658778205833
F1 Weighted: 0.9921286045231726
```

Hyperparameter Tuning

- Hyperparameter tuning was performed to optimize model performance.
- GridSearchCV and RandomizedSearchCV were used to identify optimal parameters for models such as kNN.
- This approach improved predictive accuracy and model generalization.

```
print(grid_knn.best_params_)
print(grid_knn.best_score_)
```

```
{'n_neighbors': 7, 'p': 1, 'weights': 'distance'}
0.678528334178207
```

Business Insights

Budget Segment

The Budget Semi-Auto segment demonstrates high volume sales, appealing to cost-conscious consumers seeking value.

Premium Segment

Consumers in the Premium Front Load segment report high satisfaction, indicating strong brand loyalty and quality perception.

Promotion Opportunities

Mid-range products present ideal promotion opportunities, potentially attracting customers from both budget and premium segments.

Thank You