Customer Segmentation for Anticipating Future Purchases in E-Commerce

Table of contents

Introduction

- This project focuses on leveraging data-driven insights and predictive analytics to drive effective customer understanding and engagement in the realm of E-commerce.

- Our goal is to analyze customer behavior and develop a model for anticipating future purchases, enabling businesses to tailor their strategies and meet evolving customer expectations.

Problem Statement

Customer Understanding & Segmentation

In the dynamic E-commerce landscape, businesses struggle to effectively understand and engage their diverse customer base.

Personalisation Challenges

Tailoring marketing strategies, product offerings, and customer experiences to meet individual needs poses a significant challenge.

Future Purchase Anticipation

Anticipating future purchases becomes imperative as it empowers businesses to proactively meet customer demands, offer personalized recommendations, and foster long-term loyalty.

Resource Optimisation

Limited resources and budget constraints further exacerbate the challenge, necessitating the optimization of marketing efforts and resource allocation for maximum return on investment.

Road Map

Data Collection

A

B

Data Exploration and Pre Processing

Predictive Model Development

C

D

Personalisation Strategies

Peak into the Data!

Data Exploration

The dataset consists of 541,909 rows and 8 columns, representing invoice records containing product information sold to customers.

1. **InvoiceNo**: Unique transaction identifier for analysis.
2. **StockCode**: Product code to identify purchased items.
3. **Description:** Product description for analyzing item attributes.
4. **Quantity**: Number of units purchased in each transaction.
5. **InvoiceDate**: Date and time of transaction for temporal analysis.
6. **UnitPrice**: Price per unit of the product.
7. **CustomerID**: Unique identifier for customer segmentation.
8. **Country**: Customer's country for regional insights.

Uncovering Insights from the dataset

Data Overview

Contains purchase information from ~4000 customers over a one-year period.

Captures transactions from December 1, 2010, to December 9, 2011.

Key Variables for analysis

- CustomerID variable with 4,372 unique values enables effective segmentation and understanding of customer-specific purchasing behaviors.
- Country variable with 38 unique values allows analysis of regional preferences and variations.
- Quantity variable with 722 unique values indicates variation in quantities purchased by customers.
- InvoiceDate variable with over 23,000 unique values reflects temporal aspect and potential for trend and seasonality analysis.
- UnitPrice variable with 1,630 unique values reveals different pricing points for products.
- Analyzing customer segments helps identify high-value customers, understand their preferences, and develop personalized marketing strategies.

Data Limitations : Gaps and Considerations

- **Missing Data**: Approximately 25% of entries lack customer assignments, limiting the representativeness and generalizability of our analysis.
- **Limited time period**: The dataset covers only one year, potentially restricting the temporal applicability of our findings.
- **Limited demographic information**: Predominantly UK-based orders may limit our ability to make broader inferences about customer demographics.
- **Limited purchase context details**: Lack of information about purchase context (e.g., time, location, intent) hampers our understanding of factors driving customer purchases.

Visualization

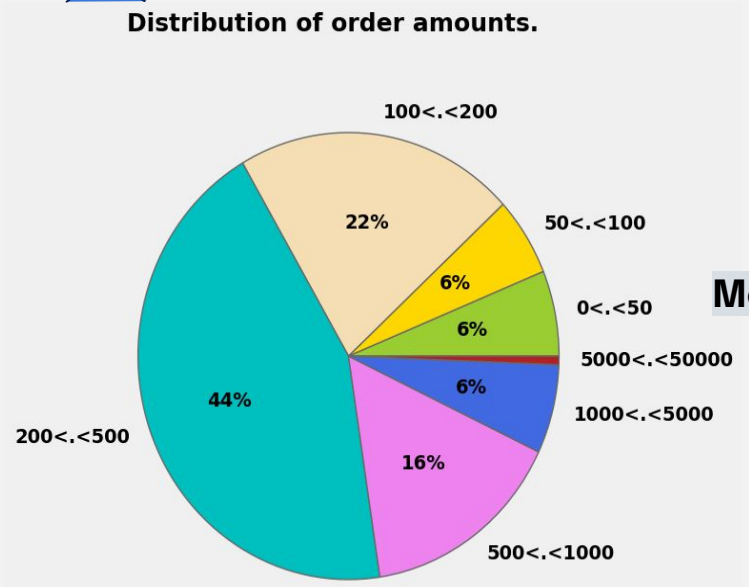
01 Proportions of Purchases & Highest occurring words

02 Quality of Clustering

03 Clustering and Frequency of certain words in Each Cluster

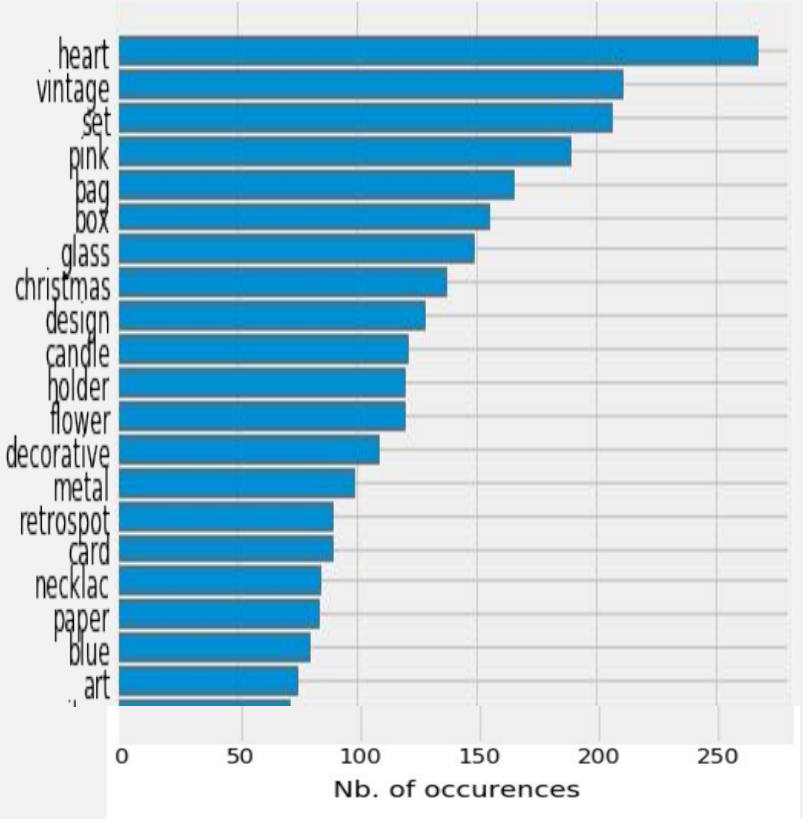
04 Variance and distribution of data Points

Data Visualization

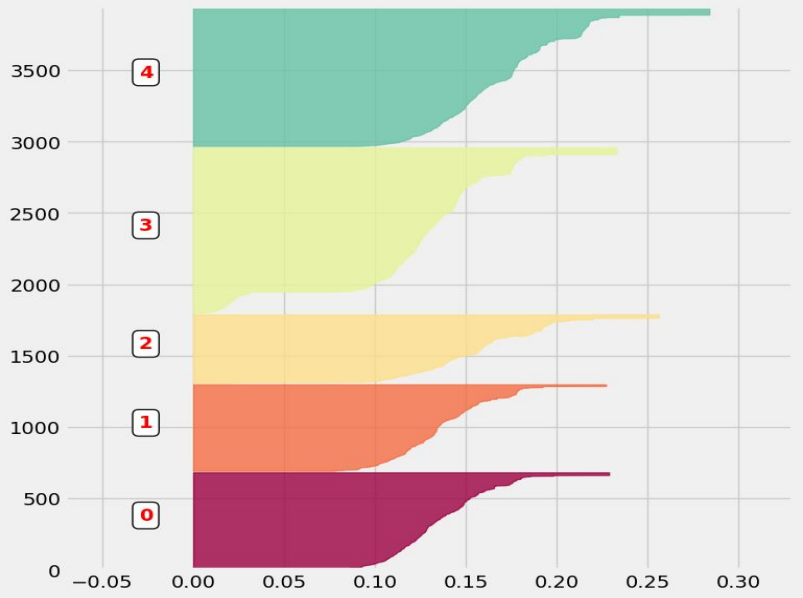


Order Amount

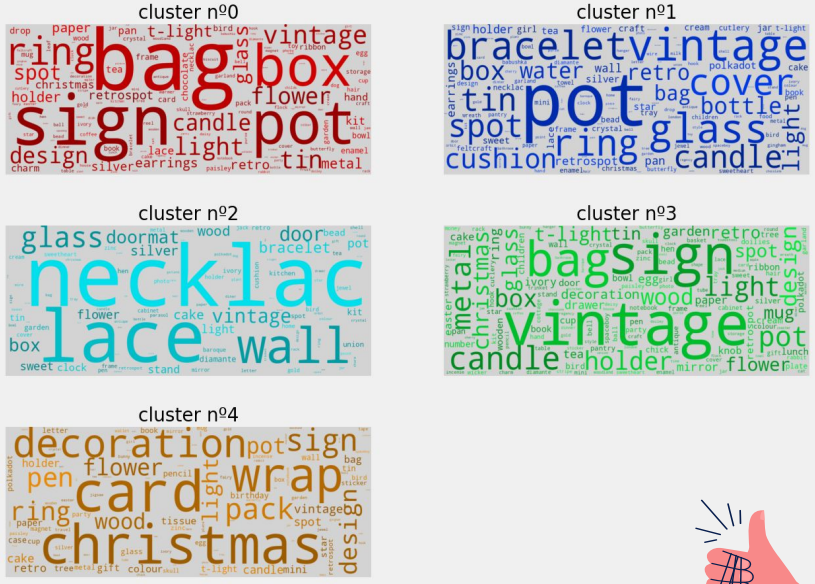
Most Used words



Data Visualization

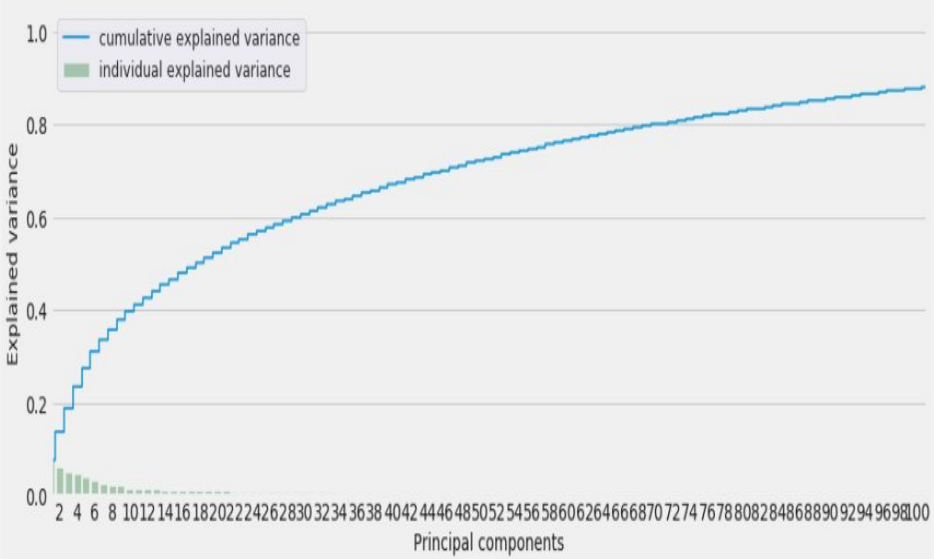


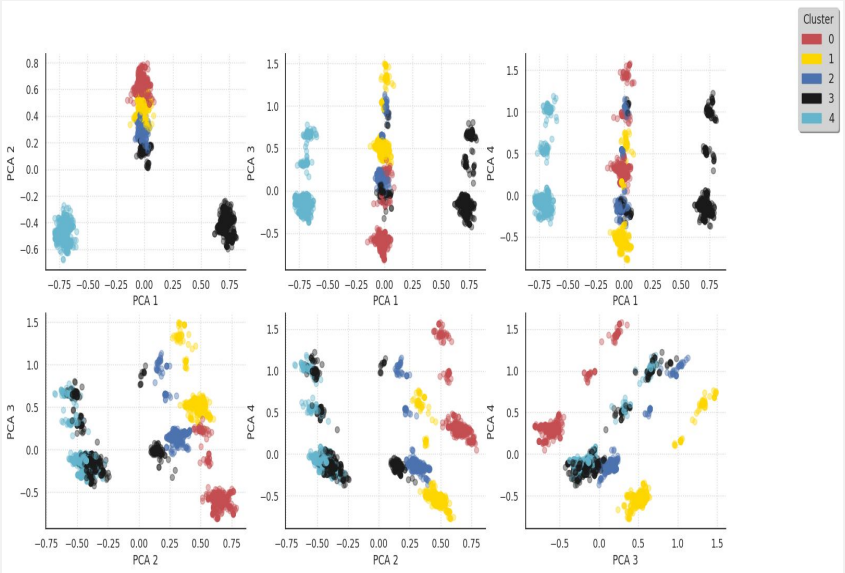
Silhouette Plot



Word Cloud

Data Visualization





Variance of Principal Component

Scatter plots(Each is the combination of 2 Principle components)

01

Model Building

Building a model is like assembling a puzzle without the box or any of the pieces—just pure fun and confusion!

Methodology and Model Architecture

Data Encoding

Two types of data encoding:

- Label encoding and
- One-hot encoding

For X and Y Variable:

- X = ['mean', 'categ_0', 'categ_1', 'categ_2', 'categ_3', 'categ_4']
- Y = ['Clusters']

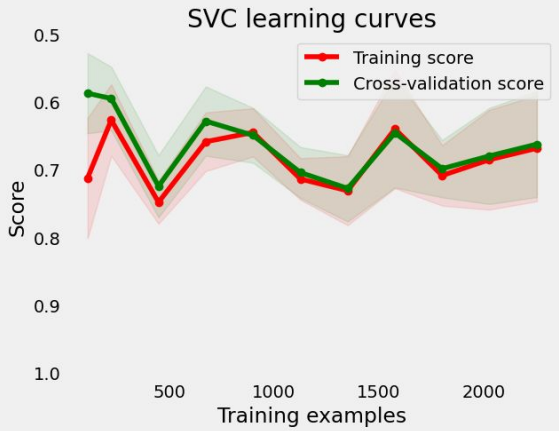
Models

```
svc = Class_Fit(clf = svm.LinearSVC)
svc.grid_search(parameters = [{'C':np.logspace(-2,2,10)}], Kfold = 5)

lr = Class_Fit(clf = linear_model.LogisticRegression)
lr.grid_search(parameters = [{'C':np.logspace(-2,2,20)}], Kfold = 5)

knn = Class_Fit(clf = neighbors.KNeighborsClassifier)
knn.grid_search(parameters = [{'n_neighbors': np.arange(1,50,1)}], Kfold = 5)

tr = Class_Fit(clf = tree.DecisionTreeClassifier)
tr.grid_search(parameters = [{'criterion':['entropy', 'gini'],'max_features':['sqrt',
'log2']}],Kfold = 5)
```



Models

```
rf = Class_Fit(clf = ensemble.RandomForestClassifier)
param_grid = {'criterion' : ['entropy', 'gini'], 'n_estimators' : [20, 40, 60, 80,
100],'max_features' :['sqrt', 'log2']}
rf.grid_search(parameters = param_grid, Kfold = 5)
```

```
ada = Class_Fit(clf = AdaBoostClassifier)
param_grid = {'n_estimators' : [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]}
ada.grid_search(parameters = param_grid, Kfold = 5)
```

```
gb = Class_Fit(clf = ensemble.GradientBoostingClassifier)
param_grid = {'n_estimators' : [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]}
```

```
mlp = MLPClassifier()
precision = precision_score(Y_test, predictions, average='macro')
```

Model	Precision
Support Vector Machine (SVM)	70.64%
Logistic Regression	79.86%
K-Nearest Neighbors (KNN)	73.90%
Decision Tree Classifier	75.04%
Random Forest Classifier	81.13%
AdaBoost Classifier	50.35%
Gradient Boosting Classifier	81.28%
MLPClassifier	72.00%

Recommendations

01 Leveraging Models with Enhanced Precision

02 Leveraging Accurate Customer Segmentation

03 Optimizing Resource Allocation

04 Enhancing Customer Retention Effort

Conclusion

05 Personalizing Marketing and Customer Experiences

06 Continuous Model Performance Monitoring

07 Ensemble Learning for Improved Performance

08 Gathering Additional Data

Conclusion

01 Optimized data modeling empowers advanced data analysis, unlocking valuable insights.

02 Precision values serve as crucial metrics, enabling accurate clustering and enhancing decision-making.

03 Precise predictions foster effective customer segmentation, facilitating personalized marketing strategies.

04 Leveraging higher precision translates to targeted marketing success, driving improved customer engagement and conversion rates.

05 The exceptional performance of Gradient Boosting algorithms in precision-driven tasks highlights their significance in delivering impactful business outcomes.