Statistical Analysis of Transactional data from a UK-based Online Retail

Wenbin Fang, Zhaoqian Xue, Michael Xie

Georgetown University DSAN

Introduction

Our study explores data mining techniques for enhancing customer-centric business intelligence for an online retailer. We aim to deepen the business's understanding of its customers to boost marketing effectiveness. Using the Recency, Frequency, and Monetary (RFM) model, we segment customers into distinct groups with k-means clustering and many other statistical learning classification methods, identifying key characteristics for each group. The findings lead to specific consumer-centric recommendations.

RFM Model & Data Preprocessing

The RFM model is a marketing analysis tool used to identify a company's best customers by measuring and analyzing three quantitative factors: Recency (R), Frequency (F), and Monetary value (M).

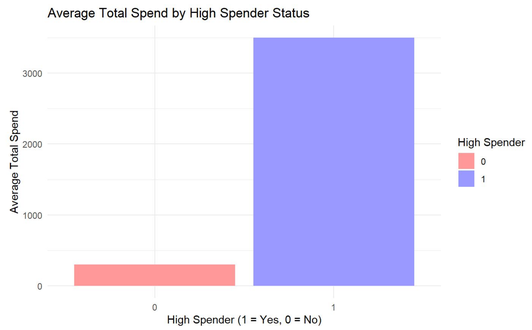
RFM
- Recency: how recently they purchased
- Frequency: how often they buy
- Monetary: how much they spend

CustomerID	TotalSpending	AverageSpending	TransactionCount	Recency
12346.0	0.0	0.0	2	325
12347.0	4310.0	23.681318681318700	182	1
12348.0	1797.24	57.97548387096770	31	74
12349.0	1757.55	24.076027397260300	73	18
12350.0	334.4	19.670588235294100	17	309
12352.0	1545.41	16.267473684210500	95	35

We clean the raw data according to the RFM model and acquire the TotalSpending(Montary), TransactionCount (Frequency), Recency.

Feature Engineering

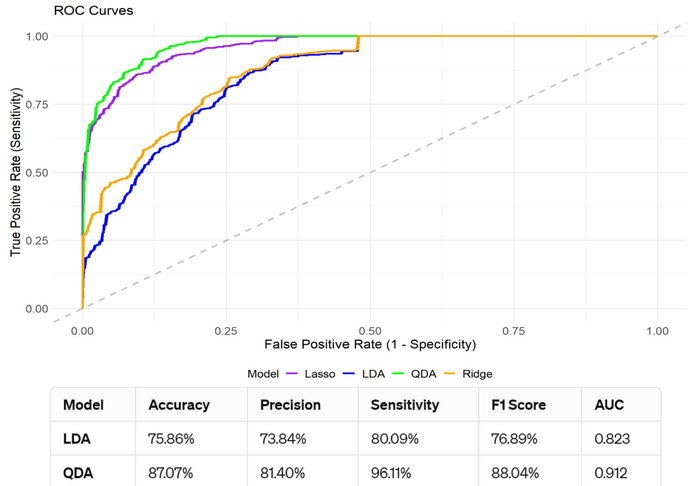
In the Online Retail dataset analysis, PCA reduced dimensions and transformed key variables like monetary value, Frequency, and Recency into uncorrelated components. The binary target **'HighSpender'** identified customers above the median spend, enhancing model accuracy for customer segmentation.



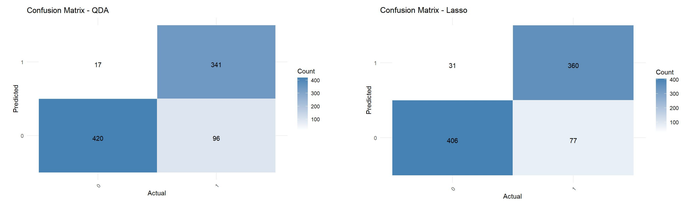
Binary Classification

In the classification task to identify "High Spenders" in the Online Retail dataset, four predictive models were employed: **Logistic Regression with Lasso and Ridge regularizations**, **Linear Discriminant Analysis (LDA)**, and **Quadratic Discriminant Analysis (QDA)**. Each model's performance was evaluated based on Accuracy, Precision, Sensitivity (Recall), and the F1 Score, illustrating their capabilities in handling imbalanced data typical in customer segmentation analyses.

Model Comparison



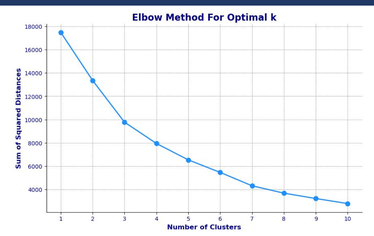
Model	Accuracy	Precision	Sensitivity	F1 Score	AUC
LDA	75.86%	73.84%	80.09%	76.89%	0.823
QDA	87.07%	81.40%	96.11%	88.04%	0.912
Lasso	87.64%	84.06%	92.91%	88.16%	0.928
Ridge	76.66%	74.53%	81.01%	77.64%	0.832



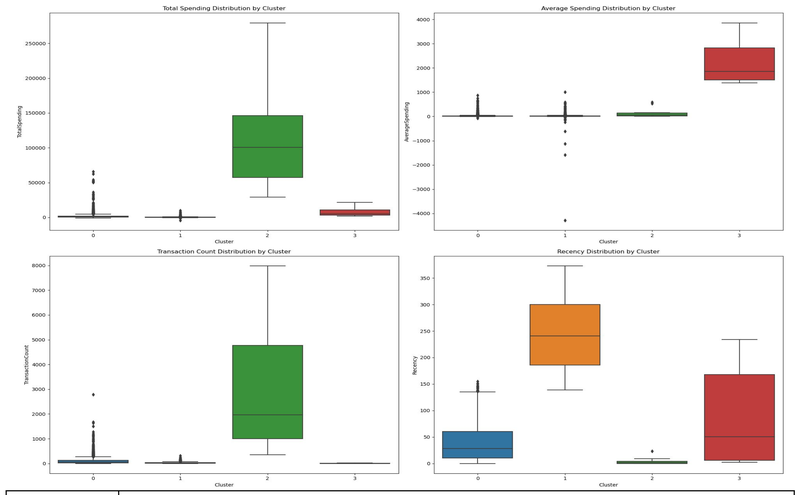
- QDA and Lasso Logistic Regression exhibited superior performance, significantly excelling in sensitivity and overall accuracy.
- LDA and Ridge models showed lower sensitivity and precision, suggesting they might struggle with the dataset's imbalanced nature.

K-means Clustering

To get more insights of segmentation of customer, we intend to group people into several groups. Thus we decide to use K-means Cluster.
By elbow method, we choose 4 as the optimal number of cluster.



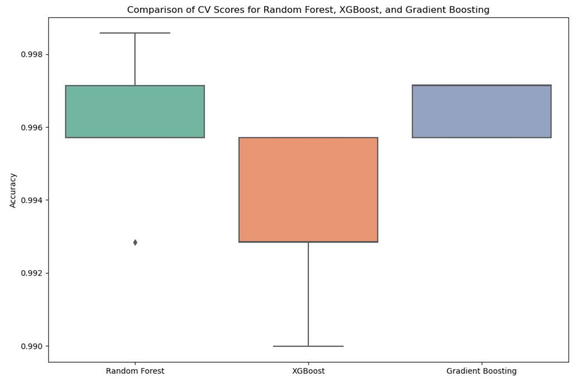
Customer Segmentation



Type 1	Shop often and spend moderate amounts, having made recent purchases.
Type 2	Infrequent, low-value purchases and a lack of recent activity in the store
Type 3	Shop frequently and spend generously, with very recent purchases
Type 4	prefer premium products and have shopped relatively recently

Ensemble ML Model

We applied three different more complicated ensemble model, which are Random Forest XGBoost and Gradient Boost, and perform the 5 fold cross-validation. We conclude that Random Forest has more stable and accurate performance with over 99.6% accuracy.



Conclusion

To wrap all the results up, we conducted two experiments on the online retail customer segmentation.
The first one is a binary classification, we split data into 2 classes by monetary(total spend), and employed 4 different methods. The logistic regression model with LASSO regularization is the optimal one.
The second we attempted to get more insights of customer. So we grouped them into 4 parts by K-means clustering and applied 3 ensemble model to do the classification. And Random Forest is the most stable and accurate.
From our research, online retailers can label their customers out of 4 types(or more), develop their recommendation algorithm respectively.