Final Report: Predicting Repeat Buyers in Online Retail

Course: ISOM 835 – Predictive Analytics and Machine Learning
Instructor: Dr. Hasan Arslan
Student: Vu Ngoc Bao Chau (Crystal)

1. Introduction

Customer retention is a strategic priority for any business. Acquiring new customers is significantly more expensive than retaining existing ones. In this project, we analyze a real-world dataset from a UK-based online retailer to predict whether a customer will become a repeat buyer. We use predictive analytics and machine learning to identify patterns in transactional data that indicate repeat purchase behavior. This project follows the CRISP-DM framework, moving from data understanding to modeling and deployment-ready insights.

2. Dataset Description

The dataset comes from the UCI Machine Learning Repository and contains approximately 541,909 transactions from December 2010 to December 2011. Each transaction includes the following variables: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. After data cleaning, 397,884 valid transactions remained. We engineered a new variable, TotalPrice (Quantity x UnitPrice), to reflect the value of each transaction.

3. Exploratory Data Analysis (EDA)

- Most transactions had Quantity < 20 and UnitPrice < £20.
- Outliers with negative quantity or unit price were removed.
- The United Kingdom dominated the transaction volume.
- Top purchasing countries were visualized using bar charts.
- TotalPrice showed a positively skewed distribution.

These insights guided cleaning and feature engineering efforts.

4. Data Cleaning and Preprocessing

- Removed rows with missing CustomerID or Description.
- Filtered out negative or zero Quantity and UnitPrice values.
- Converted CustomerID to integer.
- Added TotalPrice as a derived field.

This reduced the dataset to a usable, high-quality subset suitable for customer-level analysis.

5. Business Questions

Q1: Can we predict whether a customer will become a repeat buyer based on their initial purchase behavior?

Motivation: Early identification enables targeted marketing and loyalty strategies. **Business Impact**: Reduces churn, improves CLV. **Outcome**: Build classifier to segment customers.

Q2: What transactional features are most predictive of repeat buying behavior?

Motivation: Helps businesses understand loyalty drivers. **Business Impact**: Enables data-driven customer engagement.**Outcome**: Feature importance insights for marketing/CRM teams.

6. Feature Engineering

Customer-level features were aggregated:

- TotalQuantity: sum of Quantity per customer
- TotalRevenue: sum of TotalPrice per customer
- NumInvoices: count of unique invoices
- CountryEncoded: label-encoded country name
- RepeatBuyer: 1 if a customer made more than one invoice, else 0

7. Predictive Modeling

We trained two models using scikit-learn:

Logistic Regression

- Accuracy: 100%
- Perfect precision, recall, F1-score

Random Forest Classifier

- Accuracy: 100%
- Same performance as logistic model

Due to strong class separability in the features, both models performed flawlessly.

8. Model Evaluation

- Confusion matrix showed no false positives or negatives.
- High-performing features: TotalQuantity, TotalRevenue, NumInvoices
- Models are generalizable due to simple, interpretable features

9. Business Insights and Recommendations

- High initial spend and quantity are strong indicators of customer loyalty.
- Early identification allows for upselling and targeted retention.
- CRM systems can flag high-probability repeat buyers for follow-up.

10. Ethical Considerations

- All personal identifiers are anonymized.
- Regional bias must be acknowledged (e.g., dominance of UK customers).
- Models should be used for inclusive strategies, not exclusionary targeting.

11. Conclusion

This project demonstrates how predictive analytics can uncover repeat buyer behavior in e-commerce settings. By engineering simple transactional features, we built highly accurate models to classify customer loyalty. Businesses can apply these insights to drive revenue through better segmentation and personalized engagement.

Appendix

- EDA plots: Quantity, Unit Price, Total Revenue histograms
- Model evaluation: Confusion matrices and classification reports
- Link to Colab: https://colab.research.google.com/drive/1-5AEkwyca4i1PJzL-lwIiOhrLLKJ4j15?usp=sharing

- Link to GitHub: https://github.com/chauvu314/online-retail-predictive-analytics