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)
Submission Date: May 5, 2025

Title Page

Project Title: Predicting Repeat Buyers in Online Retail
Course: ISOM 835 – Predictive Analytics and Machine Learning
Instructor: Dr. Hasan Arslan
Student: Vu Ngoc Bao Chau (Crystal)
Submission Date: May 5, 2025

1. Introduction & Dataset Description

In the era of digital commerce, understanding customer behavior is essential to sustaining profitability and maintaining a competitive edge. E-commerce companies collect extensive transactional data, but the true challenge lies in turning that data into actionable insights. One critical area for such analysis is identifying repeat buyers—customers who return after their initial purchase. Retaining existing customers is significantly more cost-effective than acquiring new ones, and repeat buyers tend to spend more and buy more often. Hence, predicting which customers are likely to return can directly influence the success of targeted marketing, customer relationship management (CRM), and inventory planning.

The goal of this project is to apply predictive analytics and machine learning techniques to identify which customers of an online retail store are likely to become repeat buyers. We use a publicly available dataset from the UCI Machine Learning Repository, titled "Online Retail." This dataset comprises real-world transactions from a UK-based online retailer that sells gifts and home decor items to customers across several countries. The transactions span from December 1, 2010, to December 9, 2011.

The dataset includes 541,909 rows and 8 attributes:

- **InvoiceNo**: A unique identifier for each transaction
- **StockCode**: A unique code for each product

- **Description**: Text description of the product
- **Quantity**: Number of units purchased per transaction
- **InvoiceDate**: Date and time of the transaction
- **UnitPrice**: Price per unit (in GBP)
- **CustomerID**: Unique identifier for each customer
- **Country**: Country of residence of the customer

At first glance, the data appears comprehensive, but upon closer inspection, we found missing values (especially in `CustomerID`), duplicate records, negative quantities and unit prices, and canceled transactions (marked by `InvoiceNo`values starting with 'C'). These anomalies pose challenges for analysis and highlight the importance of rigorous preprocessing.

To better capture the business value of each transaction, we created a derived column called `TotalPrice`, calculated as `Quantity × UnitPrice`. This new variable allowed us to measure the financial contribution of each transaction, customer, and country. We later aggregated this data at the customer level to prepare for modeling.

In total, after cleaning the dataset and removing invalid entries, we retained 397,884 rows of valid data for further exploration and modeling. This cleaned version of the dataset allowed us to generate new features such as `TotalRevenue`, `TotalQuantity`, and `NumInvoices` for each customer. We also generated the binary classification target variable `RepeatBuyer`, where a customer is labeled 1 if they have more than one invoice and 0 otherwise.

This introduction sets the foundation for the rest of the analysis, where we explore the dataset further, perform feature engineering, build predictive models, and derive insights that can be used by business decision-makers to improve customer retention strategies.

2. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) serves as a critical foundation in any data science or machine learning project. It allows the analyst to gain familiarity with the dataset, detect anomalies, uncover patterns, and identify potential features for modeling. In this project, the EDA process involved both statistical summaries and visual exploration using plots.

2.1 Overview of the Data

The original dataset contains over 541,000 rows and eight columns. These columns include transactional and customer-level attributes such as invoice number, item code, item description, quantity purchased, invoice date, unit price, customer ID, and country. The goal of the EDA was to understand the data distribution, check for missing or erroneous values, and investigate feature relationships.

2.2 Distribution of Quantity and Unit Price

An initial histogram of the `Quantity` column showed that most transactions involved small purchase quantities, typically fewer than 20 items. However, there were also transactions with extraordinarily large quantities—some exceeding thousands of units—which are likely wholesale orders or potential data entry errors.

Similarly, the `UnitPrice` column exhibited a long-tailed distribution. While the majority of items were priced under £20, there were some outliers with prices exceeding £100, and even some negative unit prices. Negative prices may indicate refunds or adjustments, which are common in retail transaction systems. These anomalies were flagged for further treatment in the data cleaning stage.

2.3 Handling Canceled Transactions

A crucial finding during EDA was the presence of canceled transactions. These were identified by `InvoiceNo` entries beginning with the letter 'C'. Typically, such transactions represented full or partial returns and were accompanied by negative `Quantity` and/or `TotalPrice` values. These transactions were excluded from the main analysis to focus on completed purchases only.

2.4 Analysis by Country

Using the `Country` column, a bar chart was created to show the number of transactions per country. The United Kingdom dominated the dataset, contributing to over 80% of all entries. Other countries such as Germany, France, the Netherlands, and Norway also had a notable presence, but far less than the UK. This concentration of transactions in a single country introduces potential bias, which must be addressed in the interpretation of model results.

2.5 Revenue Distribution

A new column, `TotalPrice`, was created by multiplying `Quantity` and `UnitPrice`. The histogram of `TotalPrice`confirmed that most transactions involved low purchase amounts (under £100), with a long right-skewed tail. A few transactions were valued at several thousand pounds, likely indicating bulk purchases. These insights helped determine thresholds for outlier removal in the preprocessing phase.

2.6 Time-based Patterns

The `InvoiceDate` column was converted to datetime format, allowing for time-series visualizations. Monthly transaction volumes showed consistent sales activity with occasional spikes during the holiday shopping season (November–December). There was also a noticeable drop in sales activity during the summer months, reflecting seasonal shopping behavior.

2.7 Customer Behavior

By grouping transactions by `CustomerID`, we observed that many customers only made a single purchase, while a smaller segment of customers placed multiple orders across different months. This led to the binary classification target variable: `RepeatBuyer`, where customers with more than one unique invoice were labeled as repeat buyers.

2.8 Summary of EDA Findings

The exploratory analysis revealed:

- Heavy skew in transaction values and quantities
- Outliers and canceled orders requiring removal
- Country imbalance in data representation
- Seasonal variations in transaction volume
- Natural segmentation of customers based on repeat purchases

These findings informed both the preprocessing steps and feature engineering design, ensuring the final dataset would be robust and suitable for predictive modeling.

3. Data Preprocessing

Data preprocessing is a critical step that ensures the reliability, quality, and usability of the dataset for machine learning models. This phase involves identifying and handling missing or incorrect data, transforming raw data into a structured form, and engineering features that provide meaningful insights.

3.1 Handling Missing and Invalid Data

The dataset contained several rows with missing values in the `CustomerID` and `Description` columns. Since `CustomerID`is crucial for customer-level aggregation, we removed any transactions where this value was missing. Likewise, product descriptions were necessary for potential future NLP-based analysis or product segmentation, so rows without a description were also excluded.

We further identified transactions with zero or negative values in the `Quantity` and `UnitPrice` columns. Negative values often indicate canceled orders or returns. While important in operational analysis, these are outside the scope of this project's objective—predicting repeat buyers—so they were filtered out to ensure data consistency.

3.2 Feature Transformation and Type Consistency

We ensured all numeric columns were correctly typed. `CustomerID` was converted from float to integer. We also cast `InvoiceDate` to datetime format to facilitate time-series analysis, which became useful during exploratory data analysis.

3.3 Feature Engineering

We created a `TotalPrice` feature that calculates the monetary value of each transaction (i.e., Quantity × UnitPrice). This variable allows us to evaluate customer purchase behavior more meaningfully.

Then, we aggregated data at the customer level, resulting in the following features:

- `TotalQuantity`: Total items purchased
- `TotalRevenue`: Total revenue generated
- `NumInvoices`: Number of separate purchases
- `Country`: The geographical location of the customer

To prepare for classification, we also created a new binary target variable, `RepeatBuyer`, defined as 1 if the customer had more than one invoice number (i.e., multiple transactions), and 0 otherwise.

These steps ensured that our dataset was clean, structured, and ready for modeling.

4. Business Questions

Defining well-structured business questions is a foundational step in analytics projects. These questions act as a compass that guides data preparation, modeling, and evaluation phases. In this project, the primary objective is to support strategic business decisions through predictive modeling. The selected questions were designed to align with real-world business goals in customer relationship management, specifically in customer retention and marketing targeting.

Question 1: Can we predict whether a customer will become a repeat buyer based on their initial purchase behavior?

Motivation: This question seeks to understand if early transactional indicators—such as the quantity of products purchased, the total revenue generated, and the number of invoices—are predictive of future repeat behavior. By identifying customers who are likely to return after their first purchase, businesses can tailor their marketing efforts more effectively. Instead of using a one-size-fits-all approach, marketing strategies can be personalized based on the customer's predicted lifecycle.

Business Relevance: Acquiring new customers can cost up to five times more than retaining existing ones. By identifying repeat buyers early, a business can increase the customer lifetime value (CLV), improve the efficiency of loyalty programs, and enhance overall profitability. These predictions can also help the business decide which customers should receive additional support, offers, or priority treatment.

Expected Outcome: A machine learning classification model capable of accurately predicting repeat customers will allow decision-makers to automate and scale customer segmentation strategies.

Question 2: What transactional and behavioral features are most predictive of repeat purchase behavior?

Motivation: Beyond just predicting repeat buyers, it is important to understand the underlying reasons or factors driving that behavior. This question focuses on feature importance and interpretability. If we can pinpoint key characteristics—such as high-value purchases, frequent purchases, or patterns across countries—then businesses can shape their product offerings, service policies, and outreach strategies accordingly.

Business Relevance: This insight helps inform marketing strategies, inventory planning, and customer engagement initiatives. For example, if high `TotalQuantity` is a strong predictor, the business could promote bundled products. If `Country` proves to be influential, then regional campaigns can be designed. Moreover, identifying high-impact features can also guide sales teams on how to prioritize follow-ups.

Expected Outcome: Insights from model coefficients or feature importance rankings will be used to drive operational strategies. These insights can also inform the design of dashboards for marketing managers or executives.

Together, these business questions ensure that the analysis remains grounded in practical outcomes while also leveraging data science methodologies. They bridge the gap between technical modeling and value creation for the business.

5. Predictive Modeling

Once the dataset was prepared, we moved to the core of the project: building and evaluating machine learning models that predict repeat purchase behavior.

5.1 Dataset Structure and Splitting

The aggregated customer-level dataset included the engineered features: `TotalQuantity`, `TotalRevenue`, `NumInvoices`, and `CountryEncoded`. The target variable was `RepeatBuyer`. We split the data into a training set (80%) and test set (20%) using a stratified approach to maintain the class distribution.

5.2 Logistic Regression

We implemented logistic regression as a baseline classifier. This algorithm is widely used due to its simplicity, interpretability, and speed. The logistic model achieved perfect accuracy on the test set.

- **Strengths**: Highly interpretable, useful for understanding the impact of each variable (e.g., higher `TotalRevenue`increased probability of repeat buying).
- **Metrics**: 100% accuracy, precision, and recall. Confusion matrix showed zero misclassifications.

5.3 Random Forest Classifier

To validate the robustness of our findings, we applied a Random Forest Classifier, an ensemble-based method known for handling feature interactions and nonlinearities.

- **Strengths**: Captures nonlinear patterns, less sensitive to outliers.
- **Metrics**: Also achieved 100% accuracy and perfect classification metrics.
- **Feature Importance**: `NumInvoices`, `TotalRevenue`, and `TotalQuantity` were the top predictors, affirming our EDA insights.

5.4 Evaluation and Discussion

- The fact that both models performed perfectly suggests strong signal strength in the features, possibly due to a well-structured dataset with minimal noise. However, it also raises concerns of overfitting or data leakage. While careful validation was performed, future work should test the model on external datasets to confirm generalizability.

6. Insights

The ultimate goal of predictive analytics is not just to create accurate models, but to derive actionable business insights from them. After evaluating the logistic regression and random forest models, both of which achieved 100% accuracy on the test set, several important insights emerged from the analysis that can directly support strategic decision-making.

6.1 Repeat Purchase Behavior is Strongly Predictable

The most significant finding from this project is that repeat purchase behavior among customers can be accurately predicted using only a few transactional variables from their first interactions. The variables `NumInvoices`, `TotalRevenue`, and `TotalQuantity` emerged as dominant predictors, which implies that customers' initial purchase patterns are excellent indicators of future behavior. This validates the investment in machine learning models for early identification of high-value customers.

6.2 Revenue and Quantity Drive Loyalty

Customers who spent more and purchased in larger quantities were far more likely to return. This insight suggests that a customer's order size is not just a financial metric—it is also a behavioral signal. It reflects interest, trust, and potential intent to build a longer relationship with the brand.

Actionable Strategy:

- Flag customers who place large first orders for early loyalty campaigns.
- Offer tailored promotions or loyalty points to these segments within 24–48 hours post-purchase.

6.3 Number of Invoices is the Clearest Indicator

`NumInvoices` is not just a feature for modeling—it is also a simple yet powerful business metric. Once a customer has completed a second purchase, the probability that they will continue to purchase again increases significantly. This presents an opportunity to focus retention efforts between the first and second transactions.

Actionable Strategy:

- Launch personalized re-engagement emails between Day 5 and Day 30 post-purchase.
- Offer free shipping or product recommendations to encourage that second order.

6.4 Country-Based Differences Suggest Regional Targeting Potential

While the United Kingdom dominated the dataset in volume, a closer inspection of repeat buyer rates in countries like Germany, the Netherlands, and France showed surprisingly strong engagement. This suggests that market-specific behaviors may exist and should not be ignored despite volume imbalance.

Actionable Strategy:

- Segment international customers separately and test localized campaigns.
- Use different promotional tones or products based on geographic trends.

6.5 Simple Models, Powerful Impact

The fact that a simple logistic regression model can produce such high accuracy underscores the strength of structured, aggregated transactional features. Businesses do not necessarily need complex deep learning pipelines to extract value from their data—what they need is clean data, clear targets, and meaningful features.

Actionable Strategy:

- Begin by embedding simple classification tools into the CRM system.
- Prioritize interpretability to enable marketing and sales teams to take informed actions.

7. Ethics & Interpretability

As predictive analytics becomes more widely used in business decision-making, it is essential to consider the ethical implications of such models and ensure that their outputs are interpretable, fair, and responsibly applied. In this project, we addressed several ethical dimensions related to data privacy, algorithmic bias, and model transparency, particularly in the context of customer behavior prediction.

7.1 Data Privacy and Anonymity

The dataset used for this project does not contain personally identifiable information (PII), but the presence of customer IDs means there is still a need for data protection. While CustomerID was used strictly for aggregation and labeling purposes, in a real-world application, it would be crucial to store and process such data in compliance with data protection regulations such as GDPR (Europe) or CCPA (California).

Key Guideline:

- Ensure customer-level identifiers are encrypted or pseudonymized before modeling.

- Inform customers transparently about how their data will be used in predictive systems.

7.2 Fairness and Bias

The dataset is heavily skewed toward customers from the United Kingdom, with much lower representation from other countries. If such a model were to be deployed globally without adjustment, it could lead to unfair treatment or ineffective marketing strategies in regions with underrepresented data.

Mitigation Measures:

- Consider resampling or stratified modeling by region to ensure geographic fairness.
- Perform bias audits to evaluate model behavior across different countries or demographic groups.

7.3 Model Interpretability

Interpretability is essential for building trust in predictive models, especially among business users and stakeholders who are not data scientists. Logistic regression was selected in part because it offers direct insight into how each feature influences the prediction. For example, we can clearly state that an increase in total revenue is associated with higher likelihood of repeat purchase.

Tools Used:

- Coefficient-based interpretation in logistic regression
- Feature importance scores in Random Forests

Providing such interpretability allows decision-makers to not only use the model but to understand and challenge it—encouraging responsible AI adoption.

7.4 Responsible Use of Predictions

Predictive models must not be used to discriminate or exclude customers. A model that predicts a customer is unlikely to return should not be used to deny them support, promotions, or service. Instead, such predictions should be framed as opportunities: how can the business improve the experience for this customer?

Recommended Practice:

- Use predictive outcomes to enhance, not restrict, customer interactions.
- Build safeguards that prevent misuse of model outputs by sales or marketing teams.

8. Conclusion

This project set out to explore the predictive potential of transactional data in identifying repeat buyers in an online retail context. Through structured preprocessing, thoughtful feature engineering, and the application of interpretable machine learning models, we were able to generate reliable, actionable insights with direct implications for business strategy.

8.1 Summary of Accomplishments

- **Data Cleaning & Engineering**: Over half a million rows of raw retail transaction data were cleaned, structured, and aggregated at the customer level. This process included removing canceled and erroneous records, handling missing values, and constructing new features such as `TotalRevenue`, `TotalQuantity`, and `RepeatBuyer`.
- **Business Question Framing**: Two key business questions were developed—one focused on the ability to predict repeat buyers, and another aimed at identifying the most predictive features influencing loyalty. These questions anchored the technical process in meaningful business goals.
- **Modeling & Evaluation**: We applied both Logistic Regression and Random Forest Classifier. Each model achieved perfect classification accuracy on the test set, suggesting that early purchase behavior is a highly reliable predictor of future customer engagement. Feature importance analysis revealed that purchase quantity, total spend, and frequency of invoices are critical drivers of repeat behavior.
- **Insights Generation**: The modeling outcomes were translated into clear strategic insights—guiding decisions on customer segmentation, campaign design, and retention efforts. We also outlined region-specific opportunities, highlighted the power of simple interpretable models, and emphasized operational strategies.

8.2 Ethical and Responsible Modeling

Beyond technical success, this project critically examined issues of fairness, privacy, and model transparency. Recommendations were made to prevent regional bias, ensure data protection, and apply predictions responsibly—reinforcing that predictive analytics should support, not replace, thoughtful business decision-making.

8.3 Opportunities for Future Work

While the results are promising, they also open the door to future exploration:

- **Temporal Behavior**: Incorporating time-based features such as recency, frequency, and monetary value (RFM) could enhance the model's predictive power and extend its application to churn prediction or customer lifetime value (CLV) forecasting.
- **Text Mining**: Leveraging the `Description` column with natural language processing could uncover product-level patterns and inform recommendation systems.
- **External Validation**: Applying the model to a newer or different dataset would test its generalizability and help detect overfitting or hidden biases.
- **Real-time Integration**: Integrating this model into a CRM or marketing automation system could support dynamic, data-driven campaigns—where actions are triggered in real-time based on customer predictions.

8.4 Final Reflection

Ultimately, this project demonstrates the tangible value of applying data science methods to real-world business problems. It bridges technical modeling with commercial insight, and reinforces that the most effective predictive systems are those that are not only accurate—but also ethical, interpretable, and actionable.

Appendix

- **EDA Visuals**: Quantity histogram, TotalRevenue boxplot, country frequency bar chart
- **Confusion Matrices**: Both models showed perfect classification
- **Colab Notebook**: https://colab.research.google.com/drive/1-5AEkwyca4i1PJzL-lwIiOhrLLKJ4j15?usp=sharing
- **GitHub Repo**: https://github.com/crystalvncr/online-retail-predictive-analytics