UNIVERSITI TEKNOLOGI PETRONAS

DEPARTMENT OF COMPUTING

TEB 2043 DATA SCIENCE

JANUARY SEMESTER 2026

GROUP PROJECT PROPOSAL

Project Title:	Predictive Analytics for E-Commerce Customer Retention	
Category:	E-commerce	
Group Members:		
Name	**Student ID**	**Programme**
Mahesh A/L Vigneswaran	22010786	Bachelor of Computer Science (Hons)
Harresh Varma A/L Ragunathan Naidu	22011071	Bachelor of Computer Science (Hons)
Haw Jean Yung	22011237	Bachelor of Computer Science (Hons)
Paul Wong Kee Hui	22011331	Bachelor of Computer Science (Hons)
Loo Hong Sheng	22011393	Bachelor of Computer Science (Hons)

TABLE OF CONTENTS

1.0 PROJECT INTRODUCTION

1.1 BRIEF OVERVIEW

This project develops a comprehensive customer churn prediction system for e-commerce businesses using machine learning techniques. By analysing transactional data from the UCI Online Retail II dataset, we will identify customers at risk of churning and provide actionable retention strategies. The system combines predictive modelling with customer segmentation to enable personalized marketing interventions.

The project addresses a critical business challenge, which is acquiring new customers costs 5-7 times more than retaining existing ones (Reichheld & Schefter, 2000), yet the average e-commerce churn rate ranges from 70-80% annually (Statista, 2023). Our solution will help businesses proactively identify at-risk customers and implement targeted retention campaigns.

1.2 OBJECTIVES AND EXPECTED OUTCOMES

Primary Objectives:

1. **Predict Customer Churn:** Develop machine learning models to identify customers likely to stop purchasing within a defined time window.

2. Customer Segmentation: Cluster customers based on purchasing behaviour using RFM (Recency, Frequency, Monetary) analysis.

3. **Feature Engineering:** Create meaningful predictors from transactional data including temporal patterns, product preferences, and engagement metrics.

4. **Actionable Insights:** Generate business recommendations for retention strategies tailored to different customer segments.

Expected Outcomes:

- A churn prediction model achieving ≥75% accuracy with balanced precision and recall
- Clear customer segments with distinct behavioural characteristics
- Interactive dashboard displaying churn risk scores, segment distributions, and key performance indicators
- Data-driven retention strategy recommendations for each customer segment
- Comprehensive documentation of the data science workflow from raw data to deployment-ready insights

2.0 DATA DESCRIPTION

2.1 SOURCE OF THE DATASET

Dataset Name: Online Retail II

Repository: UCI Machine Learning Repository

URL: https://archive.ics.uci.edu/dataset/502/online+retail+ii

License: Creative Commons Attribution 4.0 International (CC BY 4.0) - Free for academic use

Dataset Context: This dataset contains all transactions occurring for a UK-based online retail company between 01/12/2009 and 09/12/2011. The company specializes in unique all-occasion giftware and many customers are wholesalers.

Citation: Chen, D. (2012). Online Retail II [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5CG6D.

2.2 TYPE OF DATA

Primary Type: Structured

Format: Excel files (.xlsx)

Number of Files: 2 (Year 2009-2010 and Year 2010-2011)

2.3 NUMBER OF RECORDS AND ATTRIBUTES

Total Records: 1,067,371 instances (combined from both files)

Number of Attributes: 8 core variables

Dataset Size: 44.5 MB

2.4 DESCRIPTION OF KEY VARIABLES

1. **InvoiceNo**:
    - **Data Type**: String
    - **Description**: Invoice number. Nominal. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.

2. **StockCode**:
   - **Data Type**: String
   - **Description**: Product (item) code. Nominal. A 5-digit integral number uniquely assigned to each distinct product.
3. **Description**:
   - **Data Type**: String
   - **Description**: Product (item) name. Nominal.
4. **Quantity**:
   - **Data Type**: Integer
   - **Description**: The quantities of each product (item) per transaction. Numeric.
5. **InvoiceDate**:
   - **Data Type**: DateTime
   - **Description**: Invoice date and time. Numeric. The day and time when a transaction was generated.
6. **Price**:
   - **Data Type**: Float
   - **Description**: Unit price. Numeric. Product price per unit in sterling (Â£).
7. **CustomerID**:
   - **Data Type**: Integer
   - **Description**: Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.
8. **Country**:
   - **Data Type**: String
   - **Description**: Country name. Nominal. The name of the country where a customer resides.

Derived Variables (To be Engineered):

- **TotalAmount**: Quantity × UnitPrice
- **Recency**: Days since last purchase
- **Frequency**: Number of transactions per customer
- **Monetary**: Total spending per customer
- **AverageBasketSize**: Average items per transaction
- **ChurnLabel**: Binary target variable (1 = Churned, 0 = Retained)

3.0 BACKGROUND AND PROBLEM STATEMENT

3.1 MOTIVATION FOR CHOOSING THE DATASET

Business Relevance:

The global e-commerce market was valued at approximately $16.6 trillion in 2022 and is projected to reach $70.9 trillion by 2028 (Research and Markets, 2023). However, customer retention remains a critical challenge. Research demonstrates that increasing customer retention rates by just 5% can increase profits by 25% to 95% (Bain & Company, 2020). Furthermore, the probability of selling to an existing customer is 60-70%, while the probability of selling to a new prospect is only 5-20% (Marketing Metrics, 2010).

This dataset provides real-world transactional data that mirrors the challenges faced by thousands of online retailers seeking to optimize their customer lifetime value (CLV) and reduce churn.

Technical Suitability:

The Online Retail II dataset is ideal for churn prediction because:

1. **Temporal richness:** Two years of data enables time-based feature engineering and train-test splits that respect temporal ordering. (Witten et al., 2016)
2. **Customer-level granularity:** Unique CustomerIDs allow individual-level predictions
3. **Behavioural diversity:** Mix of one-time buyers, loyal customers, and wholesalers provides varied churn patterns
4. **Real-world complexity:** Includes cancellations, missing values, and data quality issues typical of production systems

Academic Value:

This project demonstrates the complete data science pipeline: from messy real-world data to actionable business insights, showcasing skills in data cleaning, feature engineering, supervised learning, unsupervised learning, and dashboard development. The RFM framework has been extensively validated in academic literature as an effective approach for customer segmentation and churn prediction (Hughes, 1994; Khajvand et al., 2011).

3.2 PROBLEM(S) TO BE SOLVED OR QUESTION(S) TO BE ANSWERED

Business Problem:

The online retail company faces high customer acquisition costs but lacks a systematic approach to identify which customers are at risk of leaving. Customer acquisition costs (CAC) in e-commerce have increased by 222% over the past eight years (ProfitWell, 2022), making retention economics increasingly critical. Without predictive insights, marketing resources are spread inefficiently across all customers rather than focused on high-risk, high-value segments.

Specific Questions to Answer:

1. **Who will churn?** Which customers are likely to stop purchasing in the next 6 months?
2. **Why do they churn?** What behavioural patterns distinguish churn from loyal customers?
3. **Which customers matter most?** How do we prioritize retention efforts based on customer lifetime value?
4. **What interventions work?** What retention strategies are appropriate for different customer segments?

Technical Problem:

Transform transactional event data into customer-level features suitable for machine learning, handle class imbalance (churners are typically the minority class), and build interpretable models that provide actionable insights rather than black-box predictions (Verbeke et al., 2012).

3.3 IMPORTANCE OF THE ANALYSIS

Economic Impact:

- Studies show that a 10% reduction in customer churn can reduce costs by 10-15% for the average e-commerce business. (Gallo, 2014)
- Targeted retention campaigns have 3-5x higher ROI than mass marketing approaches. (Adobe, 2019)
- Early identification of at-risk high-value customers prevents significant revenue loss; even a 5% increase in customer retention can lead to profit increases of 25-95%. (Bain & Company, 2020)

Operational Benefits:

- **Marketing:** Personalized campaigns based on segment-specific behaviors increase conversion rates by up to 202%. (Monetate, 2018)
- **Inventory:** Better demand forecasting from understanding loyal vs. transient customers reduces holding cost.
- **Customer Service:** Proactive outreach to at-risk customers before they churn improves satisfaction scores.

Strategic Value:

- Shift from reactive to proactive customer relationship management (CRM) systems. (Ngai et al., 2009)
- Data-driven decision-making replacing intuition-based strategies.
- Foundation for customer lifetime value (CLV) optimization, where a 1% improvement in CLV can translate to millions in revenue for mid-sized retailers (Kumar & Reinartz, 2016).

Academic Contribution:

- Demonstrates practical application of RFM analysis in modern ML pipelines.
- Showcases handling of real-world data quality issues.
- Bridges theoretical ML concepts with business problem-solving.

References

Adobe. (2019). *Digital trends 2019: Marketing effectiveness in a multiscreen world*. Adobe Digital Insights.

Bain & Company. (2020). *Prescription for cutting costs: Loyal relationships*. Retrieved from https://www.bain.com/

Gallo, A. (2014). The value of keeping the right customers. *Harvard Business Review*. Retrieved from https://hbr.org/

Hughes, A. M. (1994). *Strategic database marketing*. Probus Publishing Company.

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, 3, 57-63.

Kumar, V., & Reinartz, W. (2016). Creating enduring customer value. *Journal of Marketing*, 80(6), 36-68.

Marketing Metrics. (2010). *Customer acquisition versus retention costs*. Unpublished manuscript.

Monetate. (2018). *The state of personalization report*. Monetate Inc.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2), 2592-2602.

ProfitWell. (2022). *The 2022 SaaS industry benchmarks report*. ProfitWell.

Reichheld, F. F., & Schefter, P. (2000). E-loyalty: Your secret weapon on the web. *Harvard Business Review*, 78(4), 105-113.

Research and Markets. (2023). *Global e-commerce market report 2023-2028*. Research and Markets Ltd.

Statista. (2023). *E-commerce customer retention and churn statistics*. Statista Research Department.

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 38(3), 2354-2364.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.