ONLINE RETAIL II - MACHINE LEARNING ANALYSIS
DETAILED TECHNICAL REPORT

This report provides a clear and detailed overview of my analysis using the Online Retail II dataset. It covers data cleaning, exploratory work, feature engineering, churn prediction models, and product recommendation rules. The aim is to show how the data was handled, how the models were built, and what business value the results bring.


```
Student: Rellika Kisyula
Date: November 22, 2025
Dataset: UCI Online Retail II (Dec 2009 - Dec 2011)
```


PROBLEM 1: CUSTOMER CHURN PREDICTION (SUPERVISED CLASSIFICATION)

- Goal: Predict which customers will NOT return within 90 days
- Best Model: Logistic Regression (ROC-AUC: 0.834)
- Churn Rate: 64.7% of customers did not return
- Key Finding: Recency is the strongest predictor of churn

PROBLEM 2: PRODUCT RECOMMENDATION SYSTEM (UNSUPERVISED ASSOCIATION RULES)

- Goal: Discover products frequently bought together
- Algorithm: Apriori (Market Basket Analysis)
- Results: 33 strong association rules with lift up to 44.6x
- Key Finding: Teacup sets show exceptional co-purchase patterns

Variable	Type	Business Meaning
Invoice	Categorical	Unique transaction ID (C prefix = cancelled)
StockCode	Categorical	Unique product identifier
Description	Text	Product name for analysis
Quantity	Numeric	Units purchased (negative = returns)
InvoiceDate	Datetime	Transaction timestamp
Price	Numeric	Price per unit in GBP
Customer ID	Categorical	Unique customer identifier
Country	Categorical	Customer's country (43 countries total)

DATASET METRICS:

- Original Records: 1,067,371 transactions
- After Cleaning: 779,425 transactions (73% retention)
- Unique Customers: 5,878 (after removing missing IDs)
- Unique Products: 4,295 distinct items
- Date Range: December 1, 2009 - December 9, 2011

- Geographic Distribution: 91% UK customers

DATASET AND CLEANING

I worked with the full Online Retail II dataset from UCI. It contains more than one million rows across two years. I combined both sheets then cleaned the data step by step:

- Removed cancelled orders.
- Removed rows with negative quantity or negative price.
- Dropped rows with missing Customer ID since those customers cannot be tracked.
- Removed missing product descriptions.
- Dropped duplicates.
- Converted InvoiceDate to datetime.
- Converted Customer ID to string.
- Standardized column names.

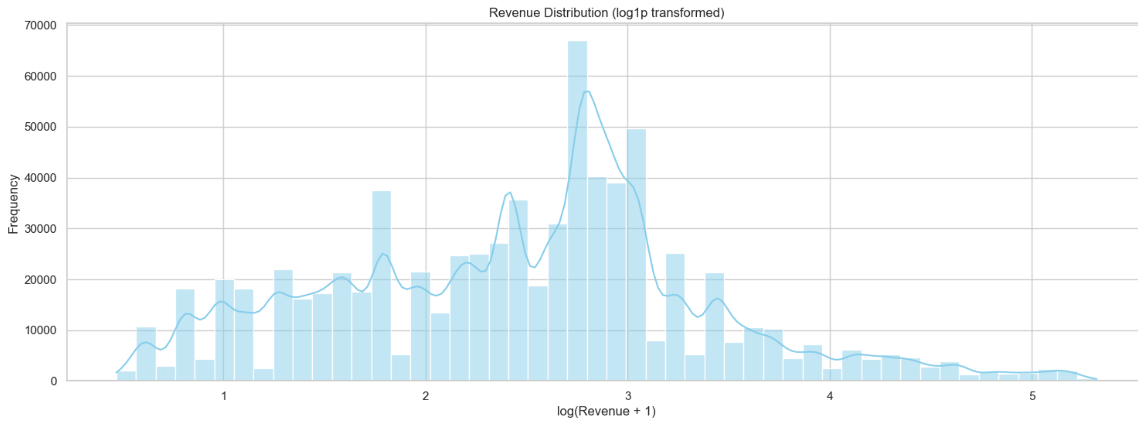
After cleaning, the final dataset had about 779 thousand valid transactions. I added the following new fields:

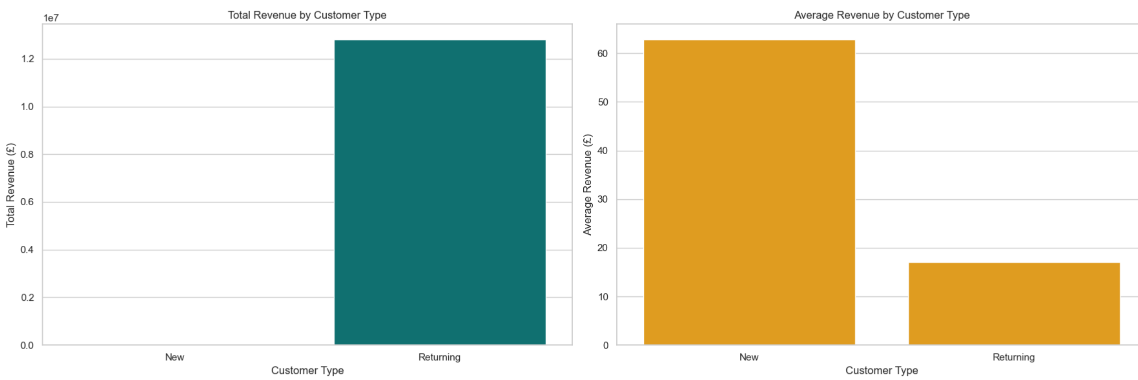
- revenue
- month
- day_of_week
- hour
- customer_type (new or returning)

EXPLORATORY ANALYSIS

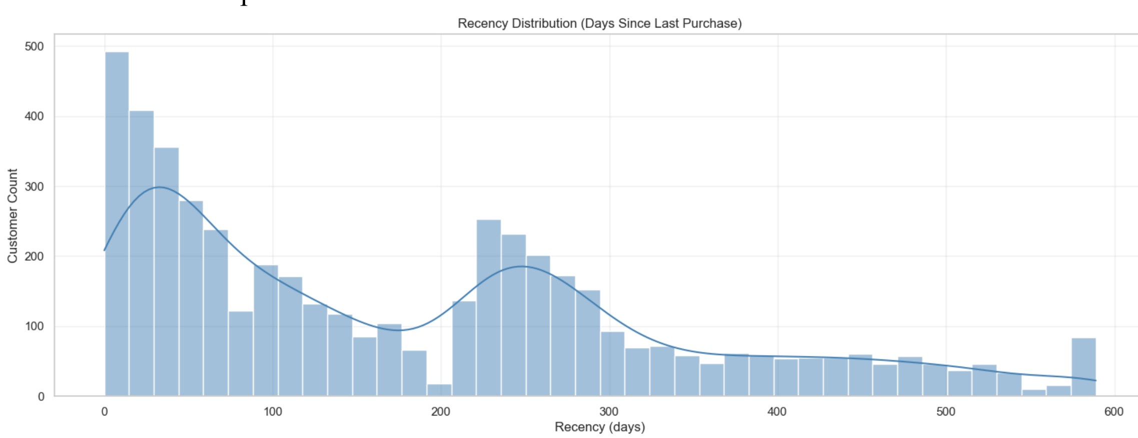
During the EDA, I explored the following fields to understand the dataset

- revenue (raw, cleaned, and log scale). The revenue was extremely right skewed. Max revenue: £168,469.60 (single transaction) and median revenue: £12.48. Long tail makes visualization difficult. I applied log scale and removed outliers (1st-99th Percentile).
- frequency of transactions
- recency (days since last purchase)
- average order value
- purchase consistency
- tenure
- total unique products per customer
- country distribution





Returning customers generate almost all of the company's revenue. The first chart shows that returning customers contribute the overwhelming majority of total revenue, while new customers account for only a very small portion. This means most revenue comes from people who come back after their first purchase.

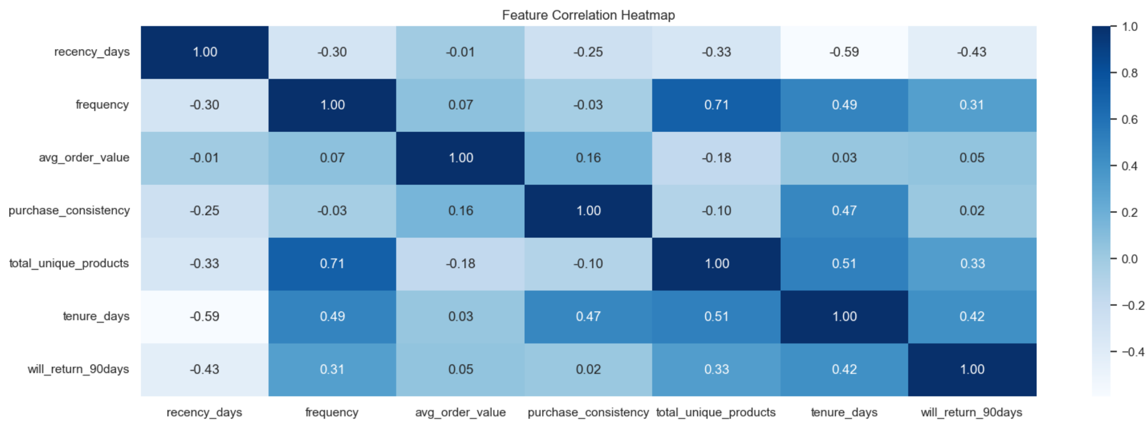


Most customers made their last purchase within the past 0 to 120 days before the cutoff date. After that, the number of active customers drops sharply. Only a small fraction of customers have recency values above 200 days, meaning most customers had some activity in the months leading up to the cutoff.

Clear patterns appeared. Recency strongly explains churn. Frequency, tenure, and average order value also help. Most customers are from the UK. Most customers buy once or twice, and only a small group buys many times.

FEATURE ENGINEERING FOR CHURN

I used a temporal split to avoid leakage. I built features only from the first 80 percent of the time range. Then I checked whether the customer returned within the next 90 days to create the churn label.



Customer level features included:

- recency_days
- frequency
- total_revenue
- avg_order_value
- total_unique_products
- purchase_consistency
- tenure_days
- country

The target variable (will_return_90days) marks whether each customer returned in the 90 day window. I ended with about five thousand customers with complete labels.

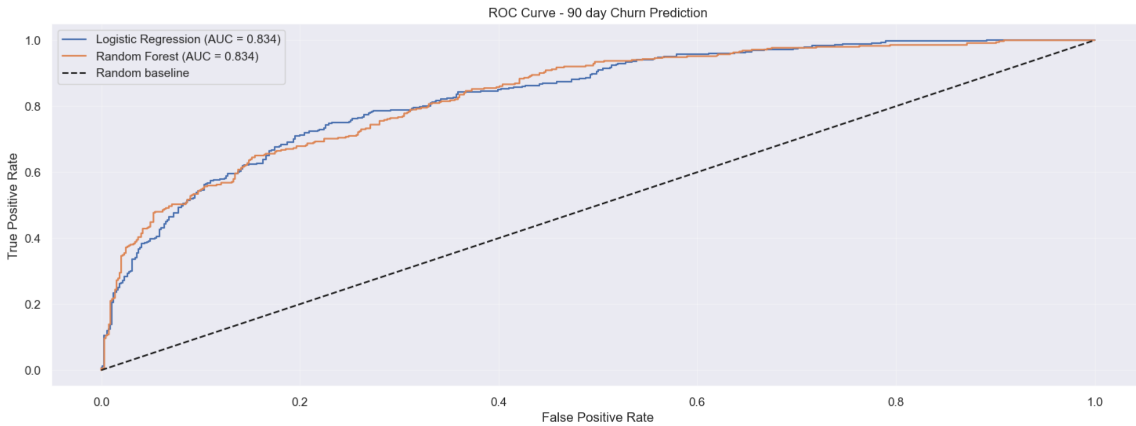
BUSINESS PROBLEM 1: CUSTOMER CHURN PREDICTION (CLASSIFICATION)

I built a full machine learning pipeline with scaling for numeric variables and one hot encoding for country.

Models I tested:

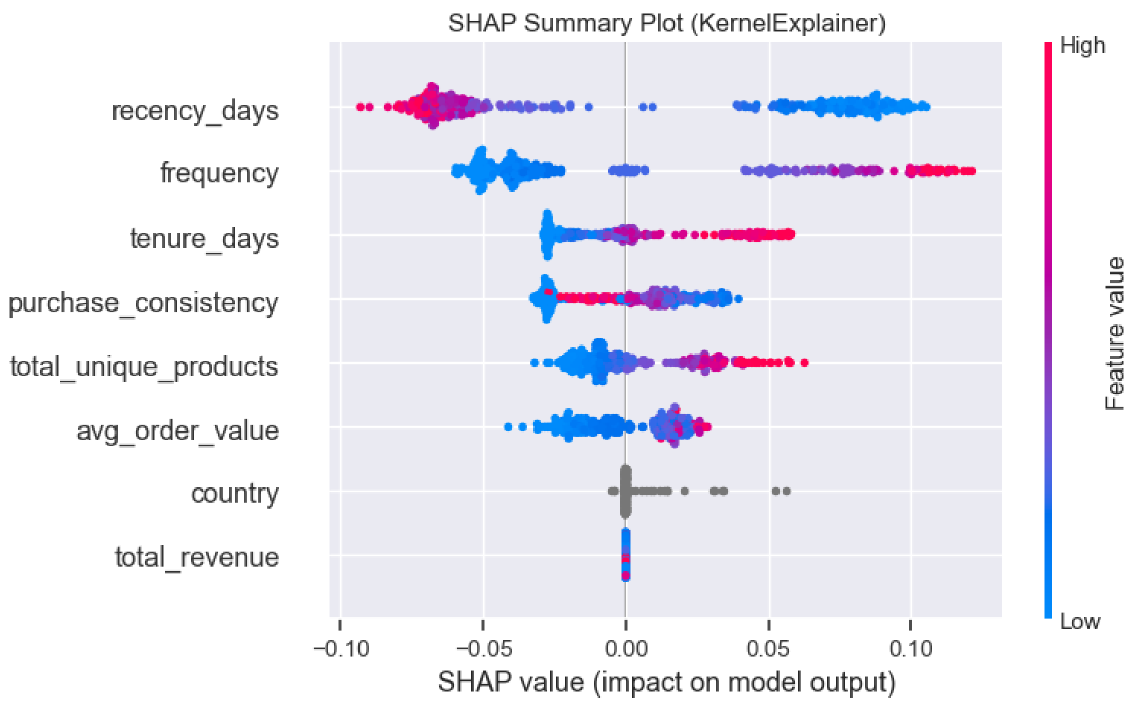
- Logistic Regression
- Random Forest
- Gradient Boosting with class weights
- Gradient Boosting with SMOTE

Model	Test AUC	Test F1	Train AUC	AUC Gap
Logistic Regression	0.834	0.630	0.813	-0.021
Random Forest	0.834	0.604	0.845	0.011
Gradient Boost (weight)	0.828	0.671	0.947	0.120
Gradient Boost (SMOTE)	0.823	0.664	0.934	0.111



Logistic Regression gave the best general test AUC at about 0.834. It did not overfit and remained stable. Random Forest had the same AUC but lower recall. Gradient Boosting caught more churners (higher recall and higher F1) but showed some overfitting.

SHAP ANALYSIS



I compared feature importance using SHAP. Recency had the highest impact. High recency pushed the model toward churn. Frequency, tenure, and average order value also helped. SHAP dependence plots showed clear non linear patterns, which explains the strong performance of tree models.

PRODUCT RECOMMENDATION SYSTEM (ASSOCIATION RULES MINING)

For product recommendations I built basket lists, filtered to typical baskets of two to twenty items, and one hot encoded them. I used the Apriori algorithm with minimum support of one percent. The model found more than two hundred frequent itemsets and thirty three strong association rules.

Top 10 Most Popular Products

Stock Code	Description	Support
85123A	WHITE HANGING HEART T-LIGHT HOLDER	9.86%
22423	REGENCY CAKESTAND 3 TIER	7.28%
85099B	JUMBO BAG RED WHITE SPOTTY	5.85%
84879	ASSORTED COLOUR BIRD ORNAMENT	5.52%
POST	POSTAGE	4.91%
47566	PARTY BUNTING	4.15%
20725	LUNCH BAG RED SPOTTY	3.52%
21212	PACK OF 72 RETRO SPOT CAKE CASES	3.43%
21232	STRAWBERRY CERAMIC TRINKET BOX	3.36%
22469	HEART OF WICKER SMALL	3.27%

The strongest rules involved the Regency Teacup sets with very high lift values. There were also strong rules for alarm clocks and trinket boxes. These rules suggest natural bundles that customers buy together.

Association Rules Mining - Find if-then patterns

#	If Customer Buys	Recommend	Support	Confidence	Lift
1	22697 (Green Regency Teacup and Saucer)	22698 (Pink Regency Teacup and Saucer)	1.20%	65.2%	43.60x
2	22698 (Pink Regency Teacup and Saucer)	22697 (Green Regency Teacup and Saucer)	1.20%	80.0%	43.60x
3	22697 (Green Regency Teacup and Saucer)	22699 (Roses Regency Teacup and Saucer)	1.37%	74.9%	36.16x
4	22699 (Roses Regency Teacup and Saucer)	22697 (Green Regency Teacup and Saucer)	1.37%	66.4%	36.16x
5	22698 (Pink Regency Teacup and Saucer)	22699 (Roses Regency Teacup and Saucer)	1.11%	74.4%	35.95x

BUSINESS VALUE

Churn Model:

The churn model can help the business find customers who might not return. It can support targeted emails or discounts. Even a small improvement in retention can lead to large gains.

Recommendation Rules:

Bundles and cross sell offers can increase the average basket size. They can also guide how products are placed on the website or in a store.
Combined Approach:

The churn scores can identify customers who need attention. The recommendation rules can provide what to offer them. This creates a full retention and cross sell strategy.

References

https://www.researchgate.net/publication/397507477_Research_on_Customer_Life-cycle_Value_Analysis_and_Refined_Operations_Based_on_UCI_Online_Retail_Data-set

https://www.kaggle.com/code/putanyn/660632067-2nd-lab-market-basket-analysis