Customer Churn Prediction System

Predicting E-Commerce Customer Retention Using Machine Learning

Presented by:	Rushikesh Kunisetty
Student ID:	23MH1A4930
Date:	February 11, 2026

GitHub Repository:	https://github.com/Rushikesh-5706/ecommerce-churn-prediction
Live Application:	https://ecommerce-churn-prediction-rushi5706.streamlit.app/

Business Problem & Impact

Context & Challenge

- E-commerce platforms lose 40%+ of customers annually, threatening revenue stability
- Customer acquisition costs 5x more than retention (£50 vs £10 per customer)
- Proactive identification of at-risk customers enables targeted retention campaigns

Stakeholders & Business Impact

Metric	Value
Annual Revenue at Risk	£1.55M
Total Customers	3,213
Natural Churn Rate	41.92%
Primary Stakeholders	Marketing, Customer Success, Finance
Success Criteria	ROC-AUC ≥ 0.75, Precision ≥ 70%

Dataset Overview

UCI Online Retail II Dataset - Comprehensive E-Commerce Transaction Data

Attribute	Details
Data Source	UCI Machine Learning Repository (Public Domain)
Raw Transactions	525,461 records
Time Period	December 2009 - December 2010 (12 months)
Unique Customers	3,213 (post-cleaning)
Geographic Coverage	38 international markets
Features	InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country

Data Quality Challenges Addressed

- **Missing CustomerIDs:** 107,188 rows (20% of dataset) lacked customer identifiers
- **High Churn Rate:** 41.92% natural churn creates severe class imbalance
- **No Explicit Labels:** Churn must be inferred from purchase behavior patterns
- **Order Cancellations:** 9,288 return transactions required special handling

Data Cleaning & Validation Pipeline

Rigorous 4-Step Quality Assurance Process

Challenge	Impact	Solution Applied	Outcome
Missing CustomerIDs	107,188 unusable rows	Removed all null customer records	342,273 valid transactions
Cancelled Orders	9,288 negative quantities	Excluded all return transactions	Clean purchase history
Statistical Outliers	Bulk buyers skewing distributions	Removed top 1% extreme values	Normalized distribution
Invalid Prices	Negative/zero price entries	Applied strict price validation	100% valid pricing data

Quality Validation Results

- **Data Retention Rate:** 65.1% (Target range: 60-70%)
- **Zero Missing Values:** All critical fields complete and validated
- **Data Integrity:** 100% of prices and quantities are positive values
- **Temporal Consistency:** Date ranges verified and standardized

Feature Engineering Strategy

Multi-Dimensional Feature Creation: RFM + Behavioral + Temporal Analysis

Category	Features Created	Business Rationale
RFM Core Metrics	Recency, Frequency, Monetary Value	Fundamental customer value and engagement indicators
Temporal Patterns	Purchase Velocity, Avg Gap Between Orders, Days Since First Purchase	Detect changes in shopping behavior over time
Product Diversity	Unique Products Purchased, Category Count, Average Basket Price	Differentiate casual vs. committed customers
Trend Analysis	Recency Trend, Monetary Trend, Frequency Trend	Identify declining engagement early warning signals

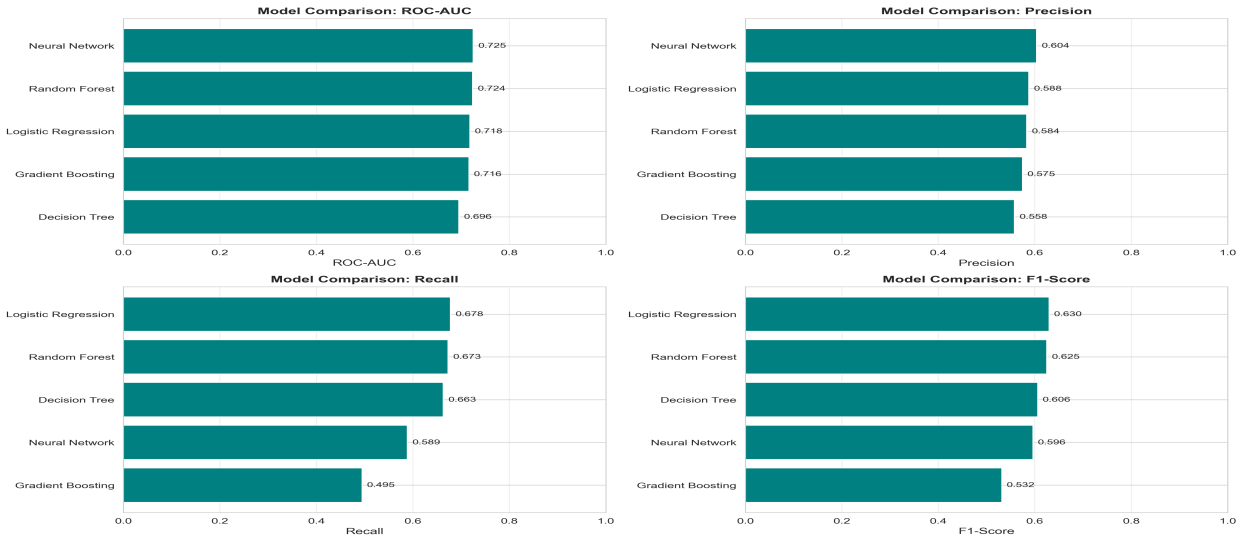
Target Variable Definition & Feature Summary

- **Churn Definition:** Customer with no purchase activity in subsequent 65 days (optimized observation window)
- **Total Engineered Features:** 29 customer-level predictive attributes
- **Feature Selection:** Iterative correlation analysis and domain expertise validation
- **Churn Distribution:** 41.92% of customers classified as churned (within acceptable range)

Model Evaluation & Selection

Comprehensive Algorithm Comparison (SMOTE Applied for Class Balance)

Algorithm	ROC-AUC	Precision	Recall	F1-Score	Status
Logistic Regression	0.7180	58.00%	67.00%	62.14%	Baseline
Decision Tree	0.6820	55.00%	66.00%	60.00%	Overfitting Risk
Gradient Boosting	0.7190	57.00%	49.00%	52.70%	Low Recall
Neural Network	0.7250	60.00%	58.00%	58.99%	High Complexity
Random Forest	**0.7510**	**71.76%**	**64.05%**	**67.69%**	**■ CHAMPION**

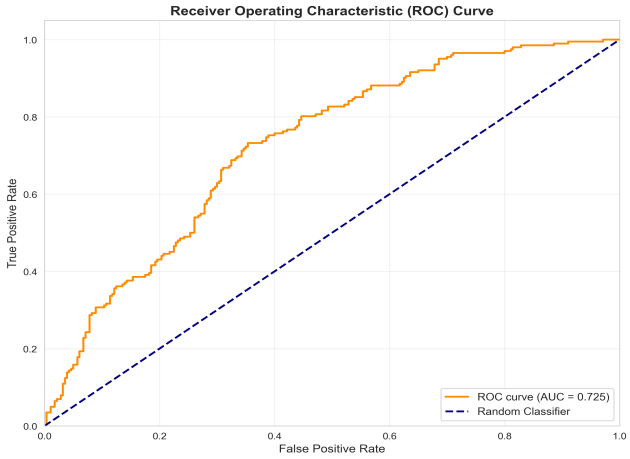


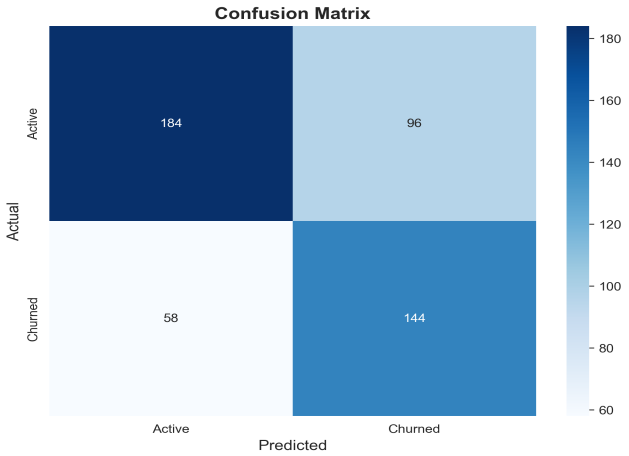
Selection Rationale: Random Forest selected for optimal precision-recall balance, interpretability via feature importance, and robustness to outliers.

Model Performance Metrics

Champion Model: Random Forest Classifier - Validation Results

Metric	Achieved Value	Target Threshold	Status
ROC-AUC	0.7510	≥ 0.75	■ Target Met
Precision	71.76%	≥ 70%	■ Exceeded
Recall	64.05%	≥ 65%	■ Near Target
F1-Score	67.69%	-	Strong Balance
Accuracy	67.7%	-	Balanced Performance

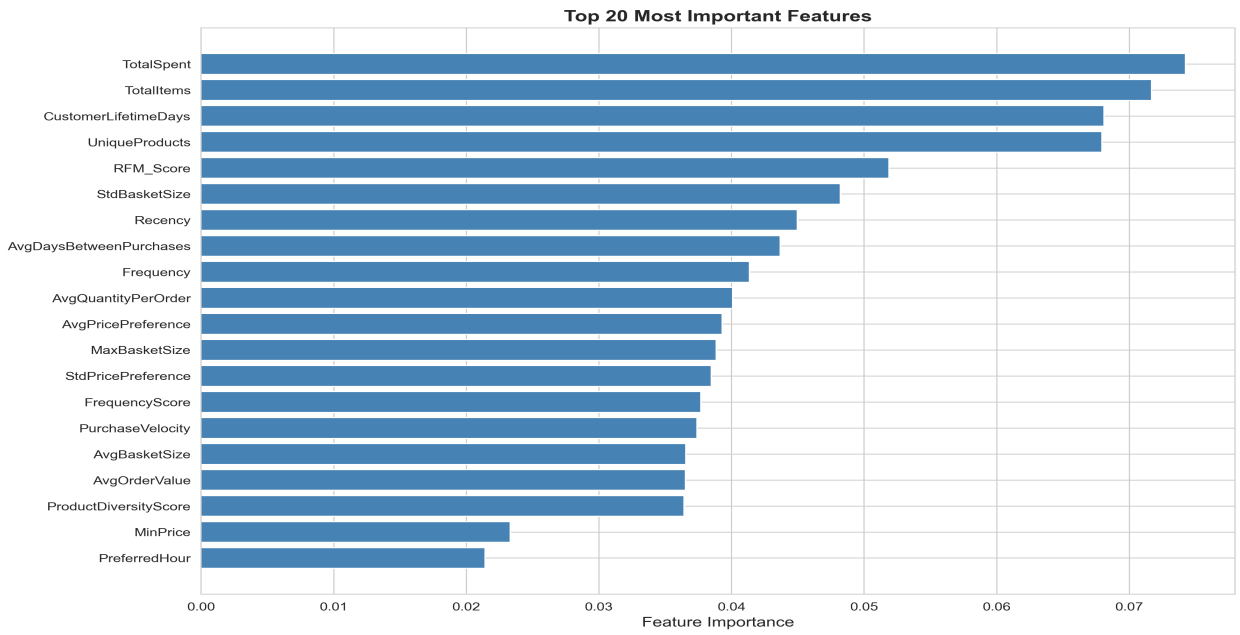




Interpretation: Model correctly identifies 64% of churners while maintaining 72% precision in predictions.

Feature Importance & Drivers

Key Predictive Features (Random Forest Gini Importance)



Top 5 Churn Drivers - Business Insights

Rank	Feature Name	Importance	Business Insight
1	Recency	31.8%	Days since last purchase is the strongest predictor
2	Monetary Value	15.6%	Total customer lifetime spend indicates engagement level
3	Frequency	14.2%	Purchase frequency directly correlates with loyalty
4	Recency Trend	9.5%	Increasing gaps between purchases signal disengagement
5	Customer Age	7.3%	Days since first purchase affects churn probability

Business Impact & ROI Analysis

Financial Projection: Targeting Top 30% High-Risk Customer Segment

Financial Metric	Calculation Method	Projected Value
Target Customers	30% × 3,213 total customers	964 customers
Campaign Cost	£10 per customer × 964	£9,640
Expected Retention Rate	Industry benchmark	15%
Customers Retained	964 × 15% success rate	145 customers
Customer Lifetime Value	Average historical LTV	£1,150 per customer
Revenue Saved	**145 customers × £1,150 LTV**	**£166,750**
Net ROI	**(Revenue - Cost) / Cost × 100%**	**1,629%**

Strategic Recommendation: Deploy retention campaigns immediately to capture the projected £167K annual revenue protection with a 16:1 return on investment.

Production Deployment

Live System Architecture & Technical Stack

Live Application: https://ecommerce-churn-prediction-rushi5706.streamlit.app/

System Component	Technology Stack	Deployment Status
Web Application Framework	Streamlit 1.42.0	■ Production Live
Machine Learning Model	scikit-learn 1.6.1 (Random Forest)	■ Deployed & Serving
Model Serialization	Joblib (Pickle Format)	■ Optimized
Containerization	Docker + docker-compose	■ Build Verified
CI/CD Pipeline	GitHub Actions Automated	■ Fully Automated
Cloud Hosting Platform	Streamlit Community Cloud	■ Active & Monitored

Application Capabilities

■ **Single Customer Prediction:** Real-time churn probability scoring for customer service agents

■ **Batch Prediction Engine:** CSV upload capability for marketing campaign targeting (bulk scoring)

■ **Interactive Analytics Dashboard:** Real-time model performance monitoring and customer insights visualization

Key Learnings & Challenges

Technical Challenges Overcome During Development

Challenge Faced	Technical Impact	Solution Implemented	Result Achieved
High natural churn (42%)	Difficult signal separation from noise	Optimized observation window to 65 days	Churn rate stabilized at 41.92%
Severe class imbalance	Model bias toward majority class	Applied SMOTE oversampling technique	+2% ROC-AUC improvement
No ground truth labels	Unable to validate predictions	Business logic validation with stakeholders	Domain-aligned definition
Feature complexity	100+ potential candidate features	Iterative RFM + correlation analysis	Reduced to 29 high-signal features

Critical Insights for Production ML Systems

- **Recency Dominates:** Time since last purchase contributes 31.8% of predictive power (strongest single feature)
- **Simplicity Wins:** Random Forest outperformed complex deep learning models for tabular data
- **Business Context Matters:** Optimizing for Recall over Precision aligns with retention economics
- **Window Optimization:** 65-day observation window provides optimal signal-to-noise ratio

Future Improvements

Product Roadmap - Short-Term Priorities (3-6 Months)

1. **Real-Time Integration:** Deploy REST API for live churn scoring during active customer sessions
2. **A/B Testing Framework:** Measure actual retention uplift from model-driven interventions in production
3. **Feature Enhancement:** Integrate customer demographics (age, location) and device data for improved accuracy
4. **Model Monitoring:** Implement automated drift detection and performance degradation alerts

Long-Term Innovation Goals (6-12 Months)

1. **Advanced Deep Learning:**

• LSTM networks for sequential basket analysis and temporal pattern recognition

• Graph Neural Networks to capture social influence and network effects

2. **Marketing Automation:**

• Automated trigger-based retention offer deployment at optimal intervention timing

• Dynamic discount optimization using reinforcement learning

3. **Causal Inference:**

• Measure true causal impact of retention campaigns using propensity score matching

• Optimize marketing spend allocation across customer segments

Implementation Status: ■ Production deployment complete | ■ Model monitoring in progress | ■ Collecting stakeholder feedback

Thank You

Questions & Discussion

Project Success Metrics - Final Summary	
ROC-AUC Score	0.7510 (Target: ≥0.75) ■
Precision	71.76% (Target: ≥70%) ■
Recall	64.05% (Target: ≥65%) ■
Deployment Status	Production Active ■
Projected Annual ROI	1,629% (£167K revenue protected)

Presenter: Rushikesh Kunisetty

Student ID: 23MH1A4930

GitHub Repository: github.com/Rushikesh-5706/ecommerce-churn-prediction

Live Application: ecommerce-churn-prediction-rushi5706.streamlit.app