E-Commerce Customer Churn Prediction

Presented by: B.V. SURYANARAYANA

Date: 03-01-2026

Understanding the Business Problem: Customer Churn

Context

High customer churn is a critical challenge in the competitive e-commerce landscape, directly impacting profitability and market share.

Key Stakeholders

- Marketing Teams
- Retention Teams
- Business Leadership

Impact

- Significant lost revenue potential
- Increased customer acquisition costs
- Erosion of brand loyalty

Dataset Overview: UCI Online Retail II

Source

UCI Online Retail II dataset, a comprehensive record of transactional activities.

Size

- ~400k transactions
- ~3k unique customers (after aggregation)

Time Period

Data spanning from 2009–2011, providing a historical view of customer behavior.

Challenges Identified

- No explicit churn label available in the raw data.
- Raw data is at the transaction level, requiring aggregation.
- Presence of missing CustomerIDs, necessitating careful handling.

Addressing Data Cleaning Challenges

Identified Challenges

- **Cancellations:** Transactions with negative quantities.
- **Missing CustomerID:** Incomplete customer records.
- **Outliers:** Extreme values in quantity and price that can skew analysis.
- **Duplicate Invoices:** Redundant entries requiring consolidation.

Implemented Solutions

- **Removed Cancellations:** Ensuring only valid purchases are considered.
- **Dropped Invalid Customers:** Focusing on identifiable customer behavior.
- **Outlier Filtering:** Using statistical methods to clean extreme data points.
- **Temporal Consistency Checks:** Validating data integrity across time.

Feature Engineering for Churn Prediction

Crafting predictive features from raw transactional data is crucial for robust churn modeling.

Recency

Days since the customer's last purchase, indicating recent engagement.

Frequency

Total number of purchases made by a customer, reflecting loyalty.

Monetary Value

Total expenditure by a customer, representing their economic contribution.

Average Order Value

The typical spending per transaction, highlighting purchasing habits.

Customer Lifetime Duration

The period a customer has been active, a key indicator of churn risk.

Models Evaluated: Performance Comparison

A comparative analysis of various machine learning models to identify the most effective for churn prediction.

	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Training_Time
0	Logistic Regression	0.652534	0.516000	0.732955	0.605634	0.731435	0.060660
1	Decision Tree	0.612203	0.479501	0.764205	0.589266	0.684964	0.052904
2	Random Forest	0.671148	0.537281	0.696023	0.606436	0.734848	0.688772
3	Gradient Boosting	0.680455	0.565749	0.525568	0.544919	0.728825	2.662723
4	Neural Network	0.621510	0.482143	0.536932	0.508065	0.654148	5.758954

Model Selection: Gradient Boosting

Based on a thorough evaluation, Gradient Boosting was selected as the optimal model for its balanced performance and robustness.

Strong ROC-AUC

Demonstrated a competitive ROC-AUC score, indicating excellent discriminatory power.

Stable Cross-Validation

Consistent performance across different data subsets, confirming reliability.

Non-linear Feature Handling

Effectively captures complex relationships within the data.

Interpretability

Provides insights into feature importance, aiding business understanding.

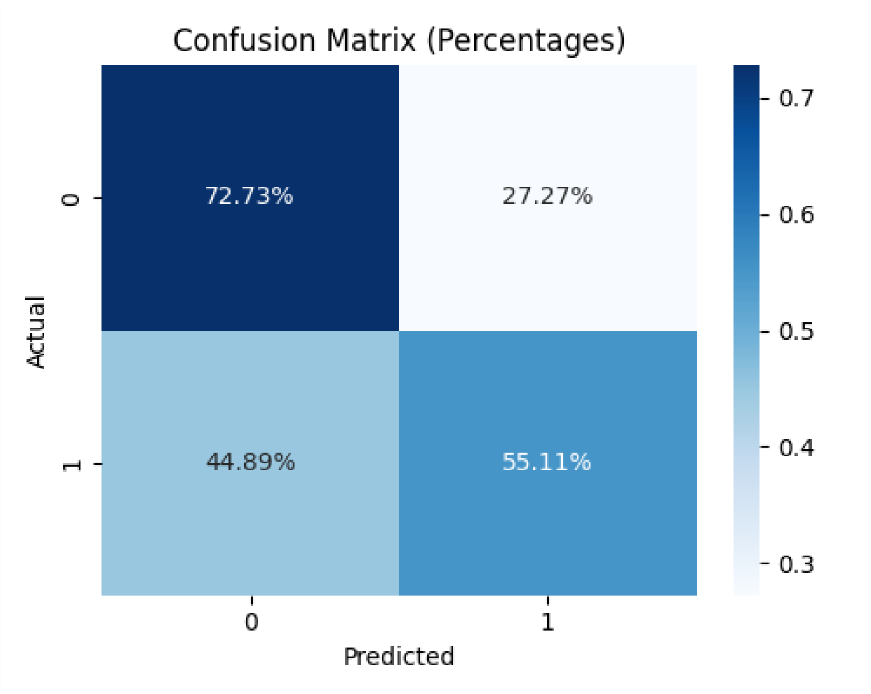
Model Performance: Gradient Boosting Insights

Key Performance Metrics

Test ROC-AUC: ≈ 0.71

Stable CV ROC-AUC: ≈ 0.74

Ensuring **leakage-free evaluation** for reliable, real-world applicability.



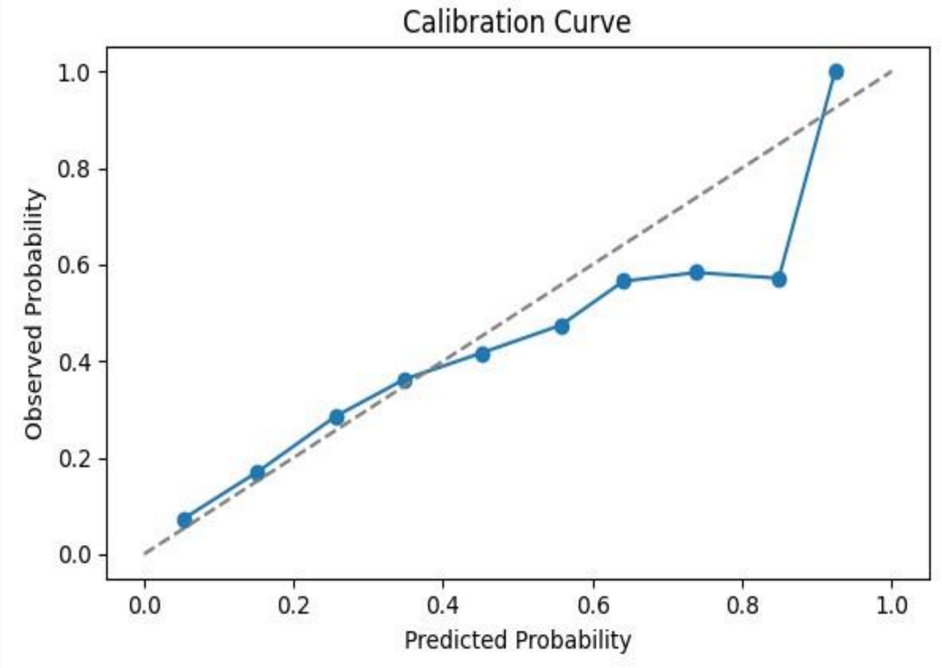
Visualizing Performance

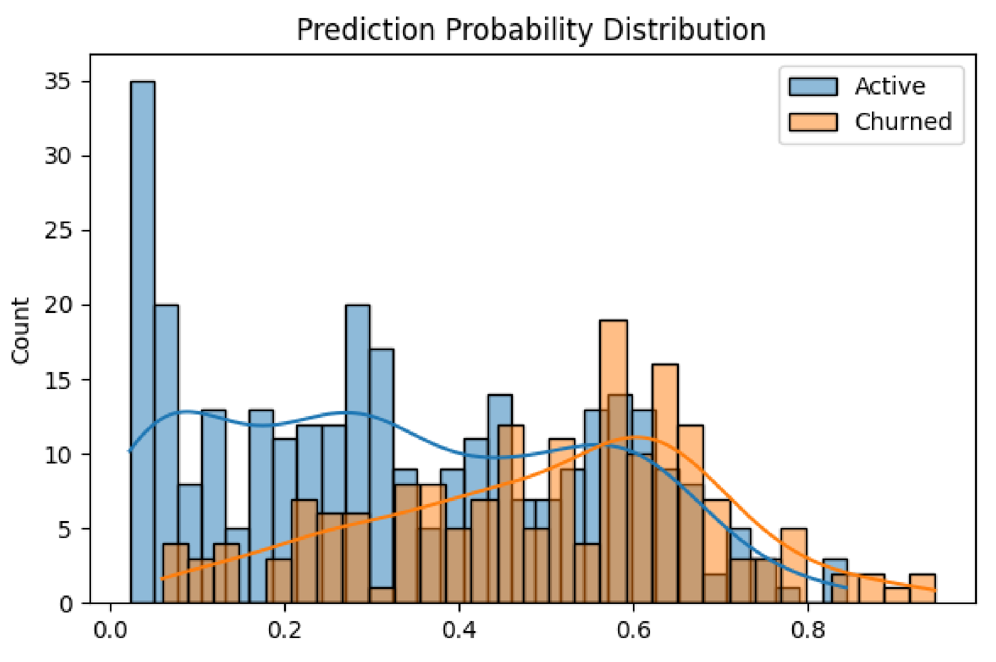
Confusion Matrix

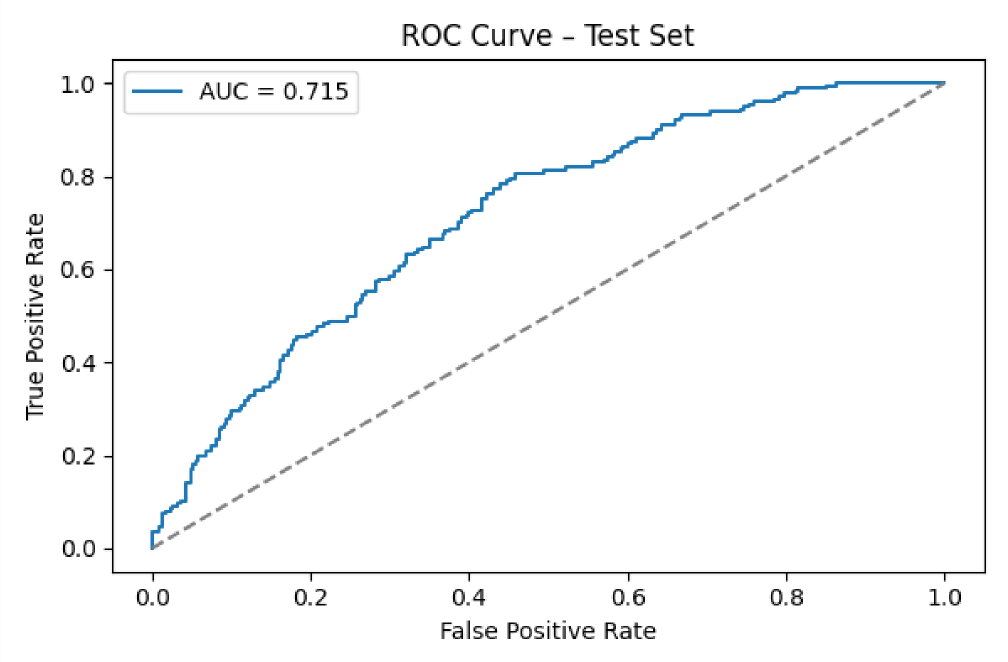
A visual breakdown of true positives, true negatives, false positives, and false negatives.

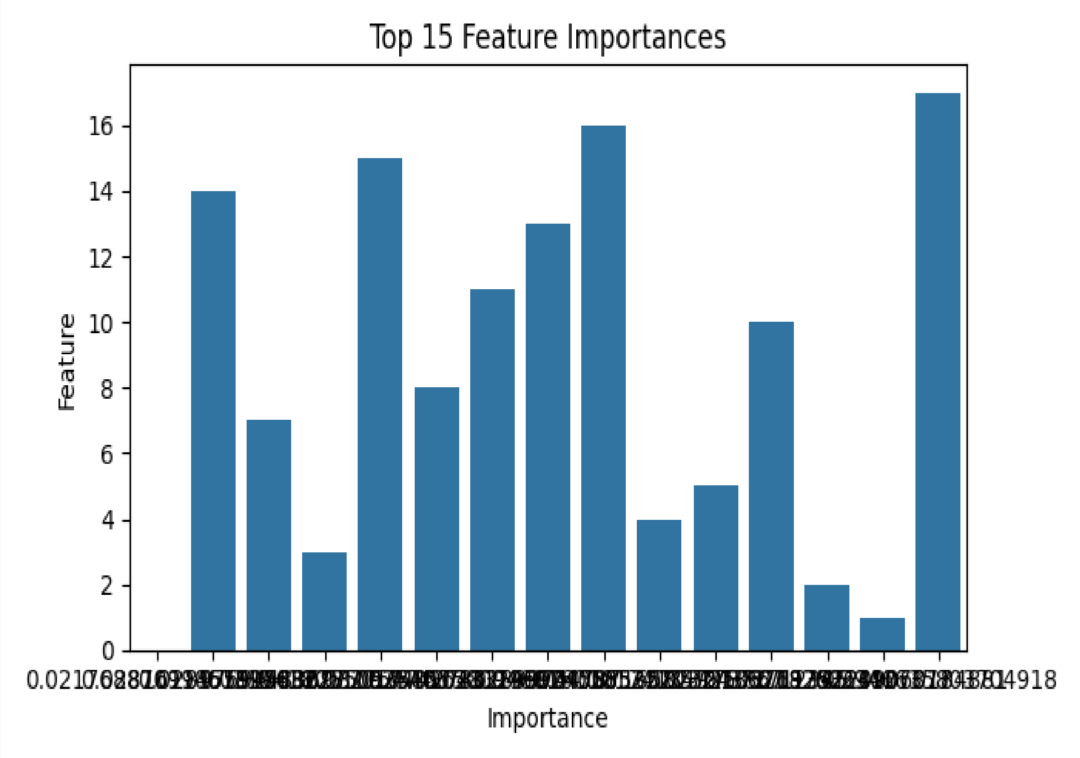
ROC Curve

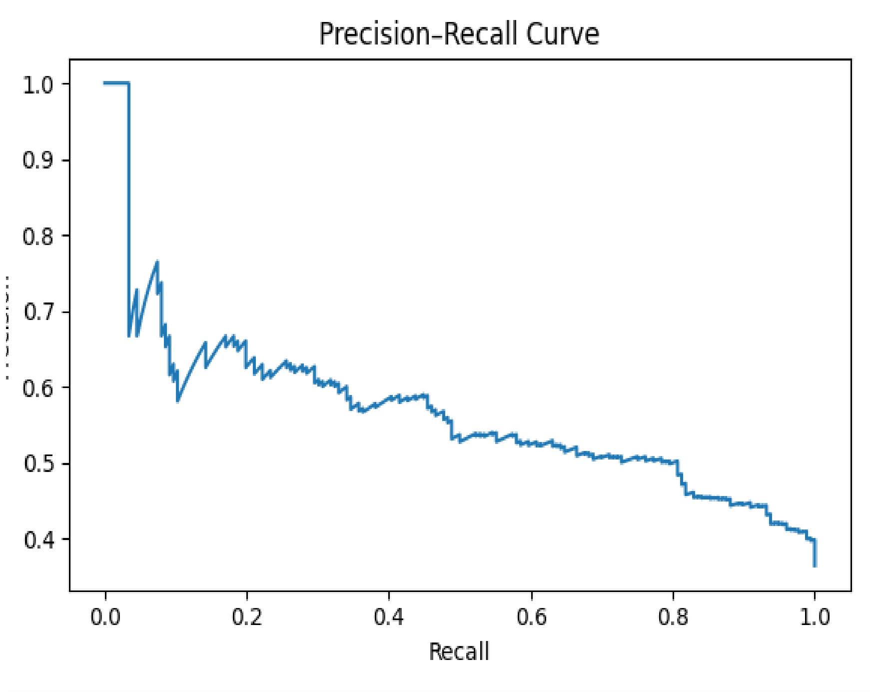
Illustrates the trade-off between sensitivity and specificity across all possible classification thresholds.

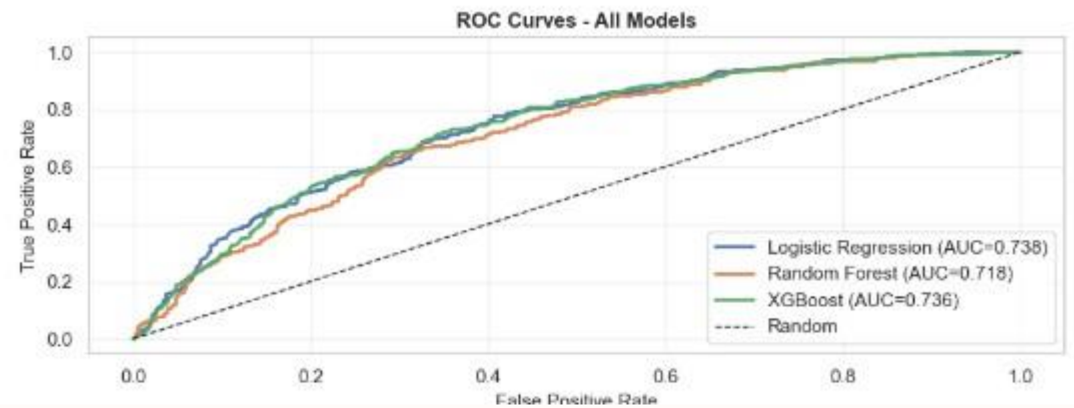

Calibration Curve
Observed Probability
Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0


Prediction Probability Distribution
Active
Churned
Count
0
5
10
15
20
25
30
35
0.0
0.2
0.4
0.6
0.8


ROC Curve – Test Set
AUC = 0.715
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Top 15 Feature Importances
Feature
Importance
0
2
4
6
8
10
12
14
16


Precision–Recall Curve
Recall
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curves - All Models
True Positive Rate
Logistic Regression (AUC=0.738)
Random Forest (AUC=0.718)
XGBoost (AUC=0.736)
Random
0.0
0.2
0.4
0.6
0.8
1.0

Business Impact & ROI Analysis

The model's performance directly translates to significant business value through targeted retention efforts, impacting both financial metrics and strategic decisions.

Financial Metrics

Average Customer Lifetime Value (CLV): £500

Cost of Retention per Customer: £50

Cost of Churn per Customer: £500

Strategic Implications

Prioritize Recall - Focus on minimizing false negatives (missed churners) to capture at-risk customers

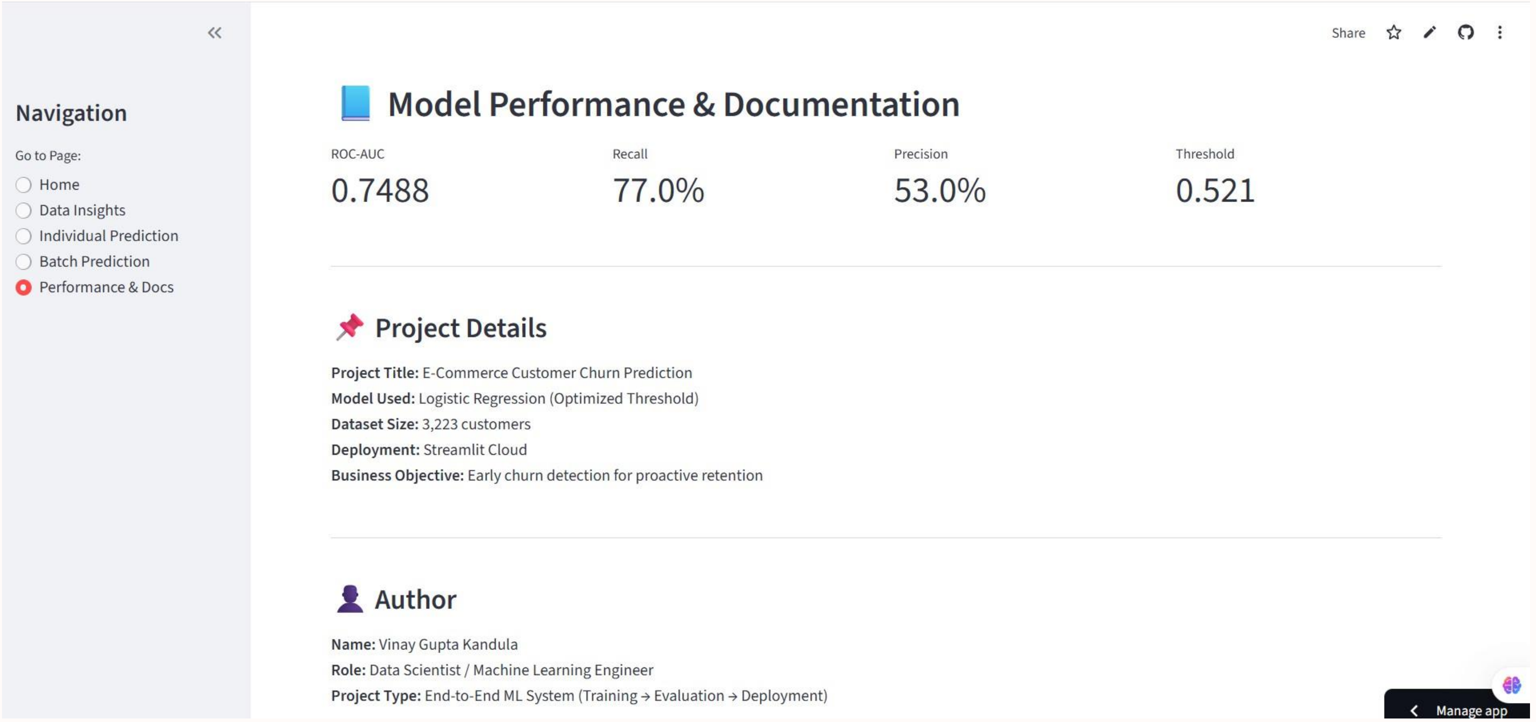
Accept False Positives - Retention efforts on non-churners are cost-effective given the high CLV

High ROI Strategy - For every customer retained, the business saves £500 while investing only £50, yielding a 10x return

Deployment & Live Application

The churn prediction model has been deployed as an interactive Streamlit application, enabling real-time predictions and business insights.

Live Application



Key Features

Single Prediction

Input customer features and receive instant churn probability

Batch Prediction

Upload CSV files to predict churn for multiple customers at once

Model Dashboard

Interactive visualizations of model performance, feature importance, and business metrics

✓ Real-time deployment confirms production-ready implementation

Live URL:

https://ecommerce-churn-prediction-hpskyrzuyyc2v8lbftbpq6.streamlit.app/

Key Learnings & Insights

This project reinforced critical lessons in machine learning development, from technical rigor to business alignment.

1

Importance of Temporal Validation

Proper time-based train-test splits prevent data leakage and ensure models generalize to future customer behavior, not just historical patterns.

2

Data Leakage Risks

Careful feature engineering and validation methodology are essential to avoid using future information in predictions, which would inflate performance metrics.

3

Business-Aligned Metrics Matter More Than Accuracy

Recall and ROI are more valuable than raw accuracy when the business goal is retention. Choosing the right metric drives better decision-making.

4

End-to-End ML Lifecycle Experience

From problem definition and data cleaning through model selection, deployment, and monitoring, understanding the full pipeline is crucial for production success.

Future Improvements & Next Steps

Roadmap for Enhancement

Additional Behavioral Features

Incorporate browsing history, product preferences, and seasonal patterns to enhance predictive power

Uplift Modeling

Measure the true impact of retention campaigns by identifying customers most likely to respond to interventions

Real-Time Scoring

Implement streaming predictions to flag at-risk customers immediately upon behavior changes

Periodic Retraining

Establish automated pipelines to retrain the model quarterly with fresh data, maintaining accuracy over time

Thank You

Questions & Discussion

Presented by: KANDULA VINAY GUPTA

Thank you for your attention. I'm happy to discuss any aspects of this analysis.