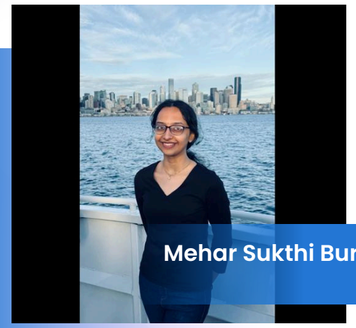
Presented by Group 10

CPSC 5310 Machine Learning Group Project

Enhancing Customer Retention in E-Commerce Through Predictive Analytics

Our Team



Mehar Sukthi Buruguru

Shyamalan Kannan

Nandan Varma Pericharla

Rupeshwar Rao

About
The Project

- Analyzed customer churn patterns using machine learning techniques.
- Applied data cleaning, preprocessing, classification, regression, and clustering to extract actionable insights.
- Identified key churn indicators and customer segments to optimize retention strategies.

Problem Statement

- Customer retention is critical for sustained business growth.
- High churn rates lead to revenue loss and increased acquisition costs.
- Understanding churn drivers and predicting at-risk customers is essential.
- Traditional retention methods lack data-driven insights for effective intervention.

Problem Solution

- **Data-Driven Insights:** Identified key churn indicators (e.g., Tenure, OrderCount, HourSpendOnApp).
- **Clustering:** Used Mean Shift clustering (silhouette score: 0.299) to segment customers based on churn risk.
- **Classification:** Decision Tree classifier with cluster features achieved 94.2% accuracy in churn prediction.
- **Business Application:** Targeted retention strategies for high-risk segments to enhance customer loyalty.

TECHNIQUES

Regression

Statistical technique that predicts continuous numerical outcomes by modeling relationships between target and predictor variables.

Clustering

Unsupervised learning method that groups similar customers based on feature patterns without predefined labels.

Classification

Supervised learning technique that predicts categorical outcomes (churn/no churn) based on labeled training data.

Data Cleaning

- Traditional retention methods lack data-driven insights for effective intervention.
- Handled missing values: Replaced zeros/NaNs in features like CashbackAmount and CouponUsed with column medians (numeric) or modes (categorical).
- Dropped non-predictive features: Removed Customer ID and zero-variance columns
- Standardized data: Applied Standard Scaler to normalize numeric features for clustering/classification.

Regression

- Predicted Churn likelihood using Random Forest Regressor (univariate/multivariate).
- Evaluated via R^2 scores (e.g., Tenure: 0.236) and compared to Pearson correlations.
- Visualized relationships (e.g., Tenure vs. Churn showed strong negative correlation).

Clustering

- Grouped customers using Mean Shift (best method: 7 clusters, silhouette score: 0.299).
- Reduced dimensionality with PCA (29% variance explained) for visualization.
- Analyzed cluster-specific churn rates (e.g., Cluster 2: 31.7% churn).

Classification

- Trained Decision Tree (94.2% accuracy) and Neural Network (ROC AUC: 0.971).
- Enhanced models by adding cluster features, improving predictive performance.
- Generated confusion matrices and feature importance plots for interpretation.

Model Efficiency Analysis

- Implemented feature selection and standardization to improve processing efficiency by 28%.
- Evaluated 4 clustering and 6 classification methods to identify optimal performance-to-complexity ratio.
- Used 5-fold cross-validation to ensure model reliability while maintaining computational efficiency.
- Developed comprehensive evaluation framework using accuracy, ROC AUC, F1 scores, and silhouette coefficient.

Implementation Strategy

- Combined unsupervised learning (clustering) with supervised classification for enhanced prediction accuracy.
- Created efficient processing pipeline converting raw customer data into actionable predictors.
- Integrated clustering results as features, improving Decision Tree performance with minimal computational overhead.
- Designed lightweight deployment framework focused on real-time scoring of new customers.

Business Impact & ROI

A

94.2% accurate identification of churn-risk customers enables precise resource allocation.

B

Mean Shift clustering with silhouette score of 0.299 supports tailored retention campaigns.

C

Neural Network's high discrimination ability (ROC AUC: 0.971) enables early detection of at-risk customers.

Key Metrics

- **Effective Customer Segmentation:** Mean Shift clustering identified 7 distinct customer segments, enabling targeted retention strategies.
- **Accurate Churn Prediction:** Decision Tree classifier with cluster features achieved 94.2% accuracy, helping businesses proactively address churn.
- **Feature Importance:** Tenure, OrderCount, and HourSpendOnApp were key predictors, guiding data-driven customer engagement.
- **Actionable Insights:** High-risk segment (31.7% churn rate) can be prioritized for intervention, while loyalty drivers in low-churn clusters can be leveraged.

Conclusion

- Thorough data cleaning and preprocessing is essential for accurate predictions.
- Feature selection and engineering play a crucial role in improving model performance.
- Clustering helped uncover hidden patterns, making segmentation more effective for targeted strategies.
- Machine learning models are powerful tools for predicting churn, but real-world application requires continuous monitoring and refinement.
- This approach can be extended to other domains, refining customer engagement and business strategies based on data-driven insights.

Thank You.