Technical Summary Report

Key Insights from Machine Learning Applications in E-Commerce

Presented By: Mehar Sukthi Buruguru, Shyamalan Kannan Rupeshwar Rao, Nandan Varma Pericharla

Focus Areas: Customer Churn Prediction, Fraud Detection, and Recommendation Systems

1. Introduction

The rapid growth of e-commerce has intensified challenges such as customer retention, fraud mitigation, and personalized user experiences. Machine learning (ML) has emerged as a transformative tool to address these issues, offering data-driven solutions for businesses to optimize operations. This report summarizes three technical papers that explore ML applications in:

- Customer Churn Prediction using hybrid ensemble models,
- Fraud Detection through systematic literature analysis,
- Product Recommendation Systems leveraging dimensionality reduction and ML algorithms.

These studies highlight advancements in accuracy, scalability, and practical implementation, while addressing shared challenges like data imbalance and computational complexity.

2. Paper Summaries

2.1. Hybrid Ensemble-Fusion Model for Customer Churn Prediction

Source: [Nature Scientific Reports]

Objective: Improve churn prediction accuracy using ensemble techniques.
Methodology:

- Combines 17 ML algorithms (e.g., SVM, Random Forest, Neural Networks) into a hybrid Ensemble Fusion framework.
- Evaluates performance via accuracy (95.35%), AUC (91%), and F1-score (96.96%).

Key Contributions:

- Outperforms standalone models (e.g., Logistic Regression, Gradient Boosting).
- Reduces overfitting through diversity in base learners.

No	ML algorithm	Precision	Recall	Accuracy	F1-score
1	Random forests[1,2]	0.94660846	0.989123	0.942994	0.967399
2	K-nearest neighbors[5]	0.8984902	0.981404	0.889289	0.938118
3	Gradient boosting classifier [6,7]	0.95816327	0.98842	0.9532	0.96306
4	Logistic regression[3]	0.87862377	0.967719	0.858086	0.921022
5	MLPClassifier(activation = 'logistic')[1]	0.94264507	0.980351	0.932193	0.961128
6	MLPClassifier(activation = 'tanh')[1]	0.93855503	0.975439	0.924392	0.956641
7	MultinomialNB classifier[8,9]	0.86323214	0.987719	0.855686	0.921289
8	BernouiliNB classifier[8,10]	0.85508551	1	0.855086	0.921883
9	GaussianNB classifier[8,9]	0.85508551	1	0.855086	0.921883
10	DecisionTreeClassifier (CART)[3]	0.95308642	0.94807	0.915692	0.950572
11	DecisionTreeClassifier (ID3)[3]	0.95149385	0.949825	0.915692	0.950658
12	SVM classifer (Linear)[10–12]	0.85508551	1	0.855086	0.921883
13	SVM classifer (Poly)[10–12]	0.92019704	0.983158	0.912691	0.950636
14	SVM classifer (RBF)[10–12]	0.92028749	0.988421	0.916892	0.953138
15	SVM classifer(sigmoid)[10–12]	0.8604878	0.928421	0.810081	0.893165
16	Adaboost classifier[6,7]	0.94765282	0.984561	0.940294	0.965755
17	ExtraTreesClassifier[5]	0.92671706	0.989474	0.924092	0.957068
18	**Our model***	**0.960088**	0.989013	**0.953533**	**0.969631**

2.2. E-Commerce Fraud Detection: A Systematic Review

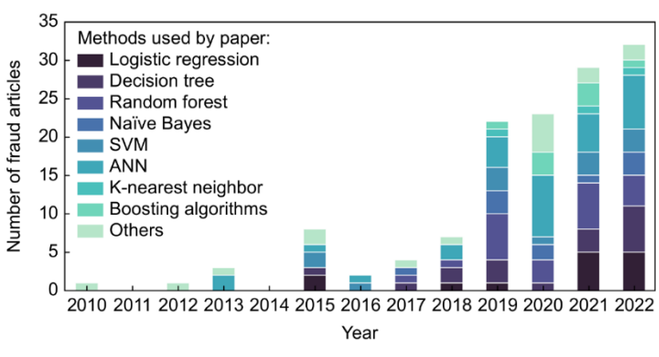
Source: [IEEE Xplore]
Objective: Identify trends and gaps in ML-driven fraud detection.
Methodology:

- PRISMA framework applied to analyse 101 studies (2013–2023).
- Highlights Artificial Neural Networks (ANNs) as a dominant approach.

Key Findings:

- Data scarcity and class imbalance limit model generalizability.
- Platform-specific fraud patterns (e.g., eBay vs. Amazon) require tailored solutions.

Visualization:



2.3. Product Recommendation System Using PCA and ML

Source: [IJCA]

Objective: Enhance recommendation accuracy with feature reduction.

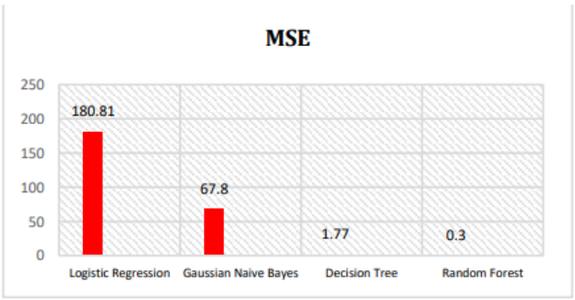
Methodology:

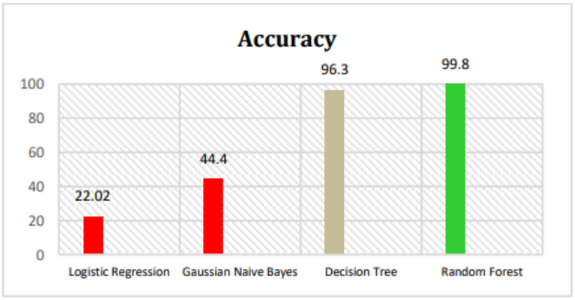
- Applies Principal Component Analysis (PCA) to reduce dimensionality.
- Tests four ML models; Random Forest (RF) achieves 99.6% accuracy.

Key Contributions:

- RF outperforms Gaussian Naive Bayes and Logistic Regression.
- PCA minimizes computational costs without sacrificing performance.
- The study demonstrates that applying Principal Component Analysis (PCA) for dimensionality reduction in product recommendation systems effectively minimizes computational costs without sacrificing performance.

Visualization:





3. Comparative Analysis

3.1. Key Techniques and Performance

Aspect	Churn Prediction	Fraud detection	Recommendation Systems
ML Algorithms	Hybrid Ensembles	ANNs, Random Forests	Random Forest, PCA
Top Accuracy	95.35%	Varies (ANN~90%)	99.6% (Random Forest)
Data Challenges	Class Imbalance	Data Scarcity, Imbalance	High Dimensionality
Preprocessing	Resampling, Normalization	Synthetic data generation	Feature Scaling, PCA

3.2. Common Themes and Challenges

Data Quality: All studies emphasize the need for balanced datasets.

Solutions: Synthetic data (fraud detection), resampling (churn prediction).

Model Complexity:

- Ensemble methods improve robustness but increase computational costs.
- PCA mitigates complexity in recommendation systems.

Real-World Scalability:

- Fraud detection requires real-time processing; churn prediction needs proactive alerts.

4. Conclusion

The three papers demonstrate ML's pivotal role in solving critical e-commerce challenges:

- **Hybrid Ensemble-Fusion Models** achieving superior accuracy in churn prediction, outperforming standalone classifiers with a **95.35% accuracy and 96.96% F1-score** [7].
- **Systematic Reviews on Fraud Detection**, emphasizing the dominance of **Artificial Neural Networks (ANNs)** while addressing challenges like class imbalance and data scarcity [16].
- **PCA-Enhanced Recommendation Systems**, where **Random Forest models with PCA** achieved **99.6% accuracy**, optimizing performance while reducing computational costs [5].

Future Directions:

- **Integration of Deep Learning for Real-Time Fraud Detection** – ANN-based models are increasingly utilized for fraud detection, offering high accuracy and adaptability in identifying fraudulent transactions in real-time
- **Personalized AI-driven Recommendations** – Combine reinforcement learning with customer behaviour analysis for dynamic and adaptive product suggestions.
- **Standardized Datasets and Evaluation Metrics** – Establish industry-wide benchmarks for fraud detection models to improve comparability and reliability.
- **Bias and Fairness in AI** – Implement techniques to detect and mitigate bias in fraud detection and credit scoring models to ensure fair decision-making.

Project Report

Enhancing Customer Retention in E-Commerce Through Predictive Analytics

Data Features and Preprocessing

Key Features Used:

Category	Features
Customer Behavior	Tenure, OrderCount, HourSpendOnApp, DaySinceLastOrder
Satisfaction Indicators	SatisfactionScore, Complain
Economic Factors	CashbackAmount, CouponUsed, OrderAmountHikeFromLastYear
Logistics	WarehouseToHome, CityTier
Technical Engagement	NumberOfDeviceRegistered, NumberOfAddress

Data Cleaning:

- Converted columns to numeric formats
- Replaced zeros with Nan in CashbackAmount & CouponUsed, filled with missing values (medians for numeric, modes for categorical)
- Standardized features for modeling, removed zero-variance columns

Prediction Models and Evaluation

Target Variable: Binary customer churn (0/1)

Regression Analysis:

- Univariate & Multivariate Random Forest Regression **($R^2 \approx 0.3$)**

Clustering Analysis:

Clustering Method	Number of Clusters	Evaluation Metrics	Best Performing
Agglomerative Clustering	3	Silhouette Score, PCA Visualization, Churn Rate Differentiation	No
K-Means Clustering	3	Silhouette Score, PCA Visualization, Churn Rate Differentiation	No
Mini-Batch K-Means	3	Silhouette Score, PCA Visualization, Churn Rate Differentiation	No
Mean Shift Clustering	**7 (auto determined)**	**Silhouette Score: 0.299**	**Yes (Best Performing)**

Cluster-Wise Churn Rates

Cluster	Number of Customers	Churn Rate	Remarks
Cluster 2	41	**31.71%** (High Risk)	**High Churn Segment**
Cluster 0	5,505	**16.89%**	Moderate Churn Rate
Clusters 3, 4, 5	50	**0%**	**Perfect Loyalty**
Cluster 6	1	**100%**	**Complete Churn**

Classification Analysis in a tabular format:

Method	Accuracy	ROC AUC	F1 Score (Churned Class)	Cluster Feature Impact
Logistic Regression	87.99%	0.855	0.518	Minimal Improvement
K-Nearest Neighbors	87.92%	0.889	0.536	No Significant Change
Decision Tree (Best Model)	**94.2%**	0.901	**0.825**	**Improved Performance**
Support Vector Machine	88.81%	0.896	0.531	Minimal Improvement
Naive Bayes	83.78%	0.800	0.526	Decreased Performance
Neural Network (Runner-up)	**94.1%**	0.971 (Best Discrimination)	**0.806**	**Improved Performance**

Evaluation Metrics Considered:

- Accuracy, ROC AUC, F1 Score (for churn class)
- Cross-validation scores, confusion matrices, comparison with and without cluster features

Conclusion & Key Takeaways

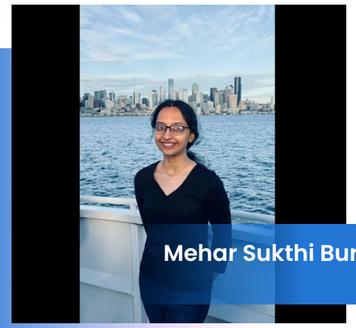
1. Customer Segmentation: Mean Shift revealed natural churn patterns, outperforming arbitrary cluster sizes.
2. Classification Effectiveness: Decision Trees balanced accuracy & interpretability, making them ideal for business use.
3. Cluster Feature Impact: Improved classification performance, especially for Neural Networks, validating a two-stage approach.
4. Business Application:
    - Target retention efforts on high-risk Cluster 2 (31.7% churn)
    - Study zero-churn clusters to identify loyalty drivers
    - Implement segment-specific retention strategies
5. Methodology Insight: The strong performance of Mean Shift **underscores the importance of data-driven clustering over predefined counts.**

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.