Predicting Customer Churn in E-Commerce Using Machine Learning

Instructor
Michael Hudson
Student
Ngan Huynh
Date: Nov 18th, 2025

Abstract. Customer retention is an important challenge in the highly competitive e-commerce industry. This project predicts customer churn using a transactional and behavioral dataset. We perform comprehensive data preprocessing, including imputation and categorical standardization, followed by exploratory data analysis to identify key drivers of churn. We then employ a Random Forest classifier, optimized with balanced class weights to handle the dataset's inherent imbalance. The proposed model achieves high predictive accuracy, with an ROC-AUC score of 0.9987, demonstrating its effectiveness in identifying at-risk customers. This methodology provides a robust.

1. Introduction

The e-commerce sector has grown exponentially, leading to intense competition. In this environment, acquiring a new customer is significantly more expensive than retaining an existing one. The phenomenon where customers cease to do business with a company is a major source of revenue loss. The ability to forecast customer attrition allows a business to shift from a reactive to a proactive retention model, engaging at-risk customers with targeted offers or improved support before they are lost.

2. Literature Review

Previous studies highlight that customer churn is influenced by factors such as tenure length, service satisfaction, complaint frequency, and purchasing behaviors. Machine learning methods, particularly ensemble models like Random Forests, have demonstrated superior performance due to their ability to capture complex, non-linear relationships. This project extends this body of work by integrating PCA visualization, class-balancing techniques, and detailed feature analysis within an e-commerce context.

3. Business Goals and Problem Definition

3.1 Business Goals

The primary business goal is to reduce revenue loss by proactively identifying and retaining customers who are at a high risk of churning.
The tactical goal is to develop a model that provides a reliable churn risk score for each customer
The strategic goal is to use this score to optimize retention marketing spend, focusing efforts on high-value, at-risk customers.

3.2 Modeling Environment

The analysis was conducted using the Python programming language within a Google

Colab environment. The core libraries utilized include pandas for data loading and manipulation, numpy for numerical operations, matplotlib and seaborn for data visualization, and scikit-learn for data preprocessing (StandardScaler), dimensionality reduction (PCA), and machine learning (RandomForestClassifier, train_test_split, and evaluation metrics).

4. Solution Methodology

4.1 Data models

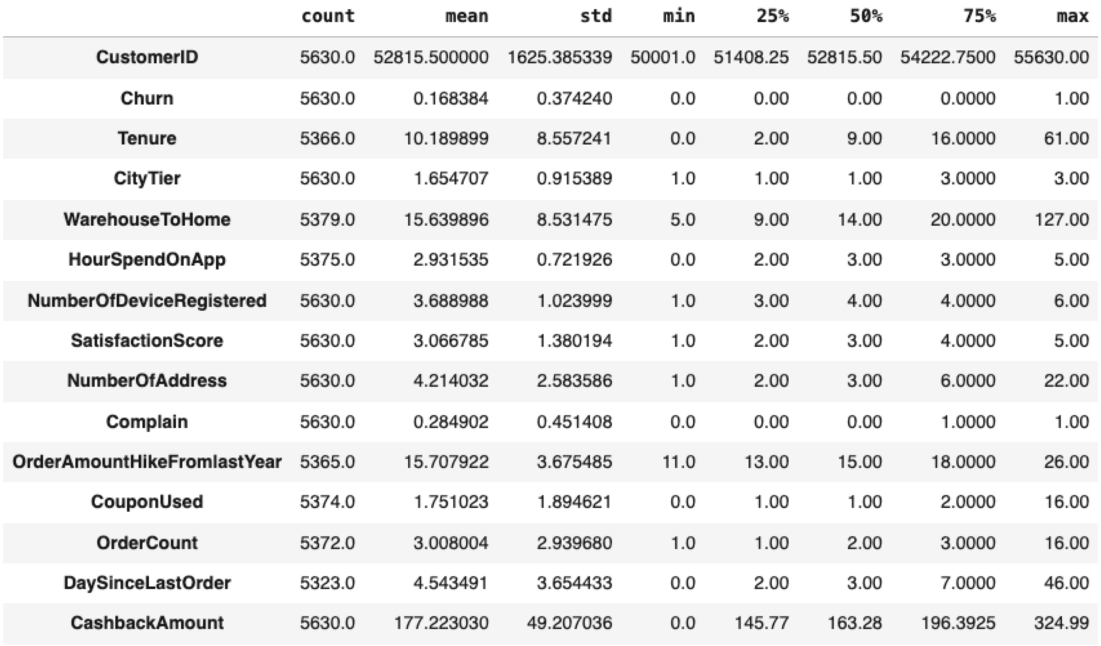
The dataset consisted of 5630 records and 20 attributes related to customer churn. The target variable is Churn, a binary flag indicating whether a customer churned (1) and a customer retained (0).

4.2 Statistical Summary and Correlation Analysis

Exploratory Data Analysis was performed to understand data distributions and inter-variable relationships.

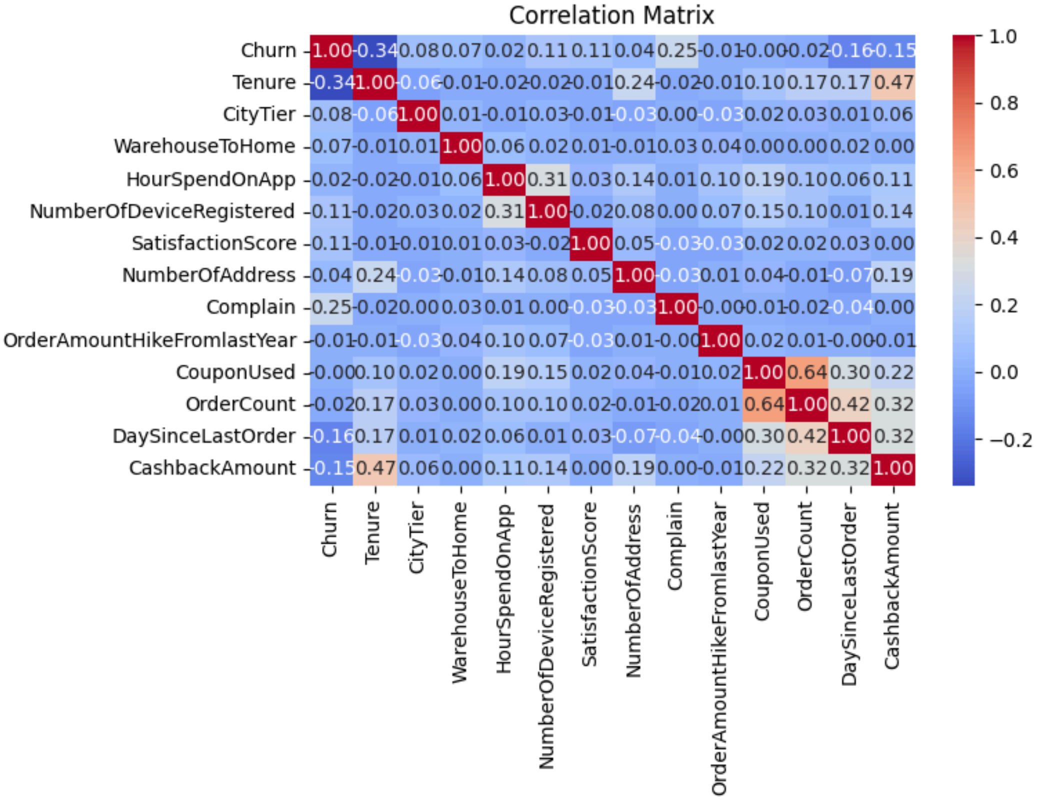
The describe().T function provided a statistical summary, revealing a significant imbalance in the target variable, with only 16.8% of customers in the 'Churn' class. This imbalance necessitates special handling during modeling.

Table 1: Statistical Summary of Numerical Features



	count	mean	std	min	25%	50%	75%	max
CustomerID	5630.0	52815.500000	1625.385339	50001.0	51408.25	52815.50	54222.7500	55630.00
Churn	5630.0	0.168384	0.374240	0.0	0.00	0.00	0.0000	1.00
Tenure	5366.0	10.189899	8.557241	0.0	2.00	9.00	16.0000	61.00
CityTier	5630.0	1.654707	0.915389	1.0	1.00	1.00	3.0000	3.00
WarehouseToHome	5379.0	15.639896	8.531475	5.0	9.00	14.00	20.0000	127.00
HourSpendOnApp	5375.0	2.931535	0.721926	0.0	2.00	3.00	3.0000	5.00
NumberOfDeviceRegistered	5630.0	3.688988	1.023999	1.0	3.00	4.00	4.0000	6.00
SatisfactionScore	5630.0	3.066785	1.380194	1.0	2.00	3.00	4.0000	5.00
NumberOfAddress	5630.0	4.214032	2.583586	1.0	2.00	3.00	6.0000	22.00
Complain	5630.0	0.284902	0.451408	0.0	0.00	0.00	1.0000	1.00
OrderAmountHikeFromlastYear	5365.0	15.707922	3.675485	11.0	13.00	15.00	18.0000	26.00
CouponUsed	5374.0	1.751023	1.894621	0.0	1.00	1.00	2.0000	16.00
OrderCount	5372.0	3.008004	2.939680	1.0	1.00	2.00	3.0000	16.00
DaySinceLastOrder	5323.0	4.543491	3.654433	0.0	2.00	3.00	7.0000	46.00
CashbackAmount	5630.0	177.223030	49.207036	0.0	145.77	163.28	196.3925	324.99

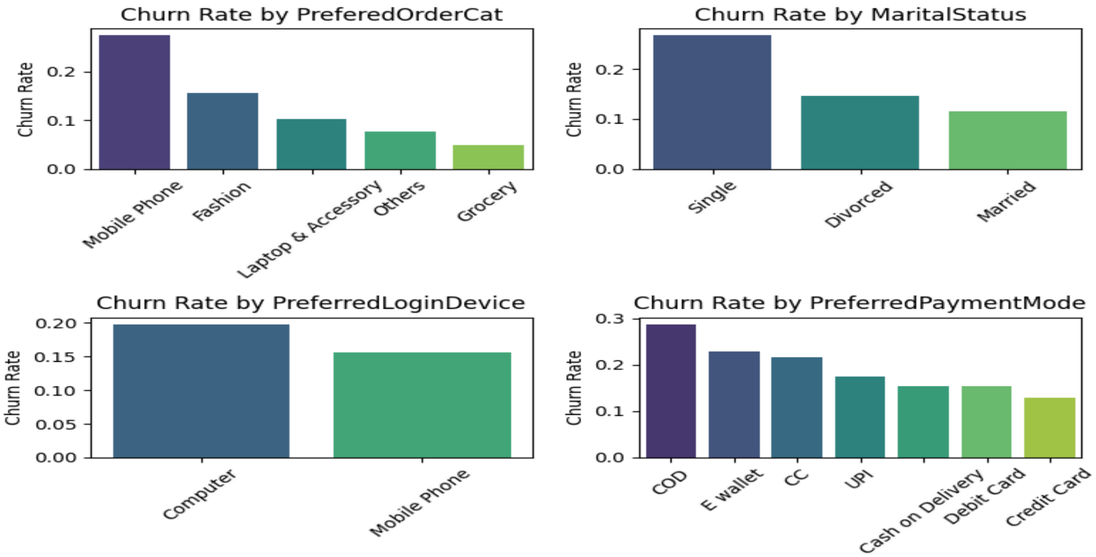
Figure 1: Correlation Matrix



	Churn	Tenure	CityTier	WarehouseToHome	HourSpendOnApp	NumberOfDeviceRegistered	SatisfactionScore	NumberOfAddress	Complain	OrderAmountHikeFromlastYear	CouponUsed	OrderCount	DaySinceLastOrder	CashbackAmount
Churn	1.00	-0.34	0.08	0.07	0.02	0.11	0.11	0.04	0.25	-0.01	-0.00	-0.02	-0.16	-0.15
Tenure	-0.34	1.00	-0.06	-0.01	-0.02	-0.02	-0.01	0.24	-0.02	-0.01	0.10	0.17	0.17	0.47
CityTier	0.08	-0.06	1.00	0.01	-0.01	0.03	-0.01	-0.03	0.00	-0.03	0.02	0.03	0.01	0.06
WarehouseToHome	0.07	-0.01	0.01	1.00	0.06	0.02	0.01	-0.01	0.03	0.04	0.00	0.00	0.02	0.00
HourSpendOnApp	0.02	-0.02	-0.01	0.06	1.00	0.31	0.03	0.14	0.01	0.10	0.19	0.10	0.06	0.11
NumberOfDeviceRegistered	0.11	-0.02	0.03	0.02	0.31	1.00	-0.02	0.08	0.00	0.07	0.15	0.10	0.01	0.14
SatisfactionScore	0.11	-0.01	-0.01	0.01	0.03	-0.02	1.00	0.05	-0.03	-0.03	0.02	0.02	0.03	0.00
NumberOfAddress	0.04	0.24	-0.03	-0.01	0.14	0.08	0.05	1.00	-0.03	0.01	0.04	-0.01	-0.07	0.19
Complain	0.25	-0.02	0.00	0.03	0.01	0.00	-0.03	-0.03	1.00	-0.00	-0.01	-0.02	-0.04	0.00
OrderAmountHikeFromlastYear	-0.01	-0.01	-0.03	0.04	0.10	0.07	-0.03	0.01	-0.00	1.00	0.02	0.01	-0.00	-0.01
CouponUsed	-0.00	0.10	0.02	0.00	0.19	0.15	0.02	0.04	-0.01	0.02	1.00	0.64	0.30	0.22
OrderCount	-0.02	0.17	0.03	0.00	0.10	0.10	0.02	-0.01	-0.02	0.01	0.64	1.00	0.42	0.32
DaySinceLastOrder	-0.16	0.17	0.01	0.02	0.06	0.01	0.03	-0.07	-0.04	-0.00	0.30	0.42	1.00	0.32
CashbackAmount	-0.15	0.47	0.06	0.00	0.11	0.14	0.00	0.19	0.00	-0.01	0.22	0.32	0.32	1.00

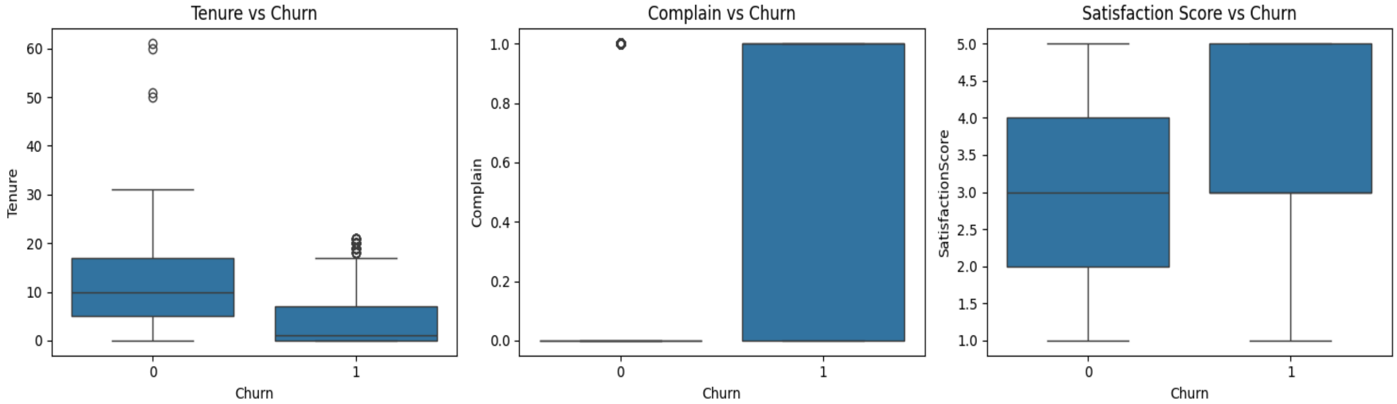
The correlation heatmap (Figure 1) shows the linear relationships between numerical features. We observed a strong negative correlation between Tenure and Churn, indicating that customers with longer tenure are less likely to churn. Conversely, Complain (whether a customer complained) showed a positive correlation with churn.

Figure 2: Churn Rate by Categorical Features



Analysis of categorical features (Figure 2) revealed several insights. The churn rate was highest for customers whose preferred order category was 'Mobile Phone'. Furthermore, 'Single' customers showed a higher propensity to churn compared to 'Married' customers.

Figure 3: Distribution of Key Features vs. Churn



As seen in the boxplots (Figure 3), the distribution of key features differs significantly between churned and retained customers. The median **Tenure** for retained customers is substantially higher than for churned customers. Notably, customers who churned had a significantly higher incidence of **Complain** (median of 1 vs 0) and a lower **SatisfactionScore.**

5. Machine Learning Models

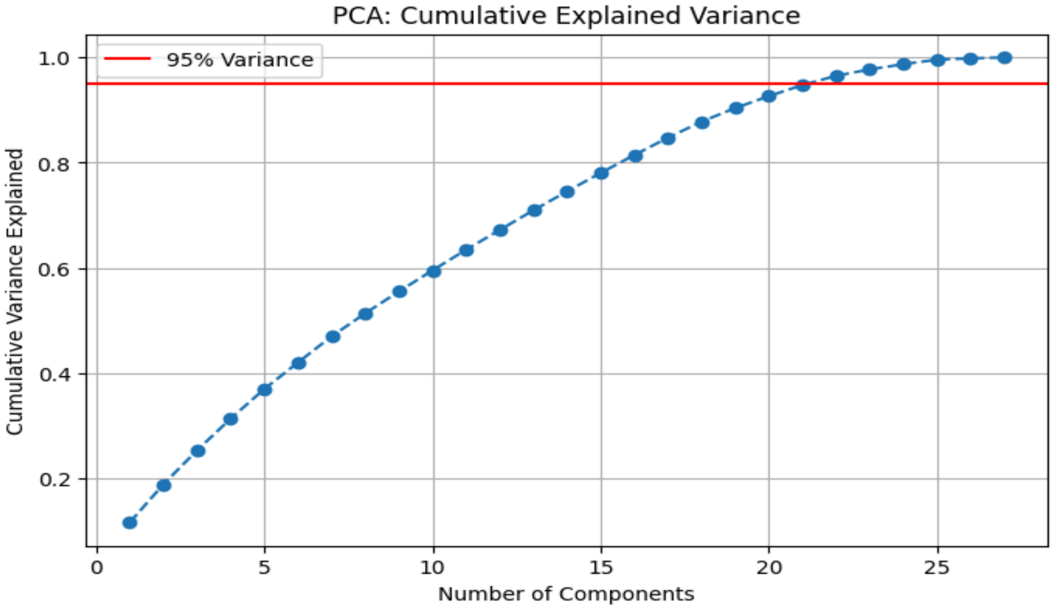
5.1 Train/Test Split and Scaling

To build and validate the model, the data were first encoded using pd.get_dummies to convert all categorical features into a numerical format.
The resulting dataset was then split into training and testing sets using an 80/20 ratio. We employed a stratified split (stratify=y) to ensure that the 16.8% churn imbalance was preserved in both sets.
This is a critical prerequisite for PCA.

5.2 Principal Component Analysis (PCA)

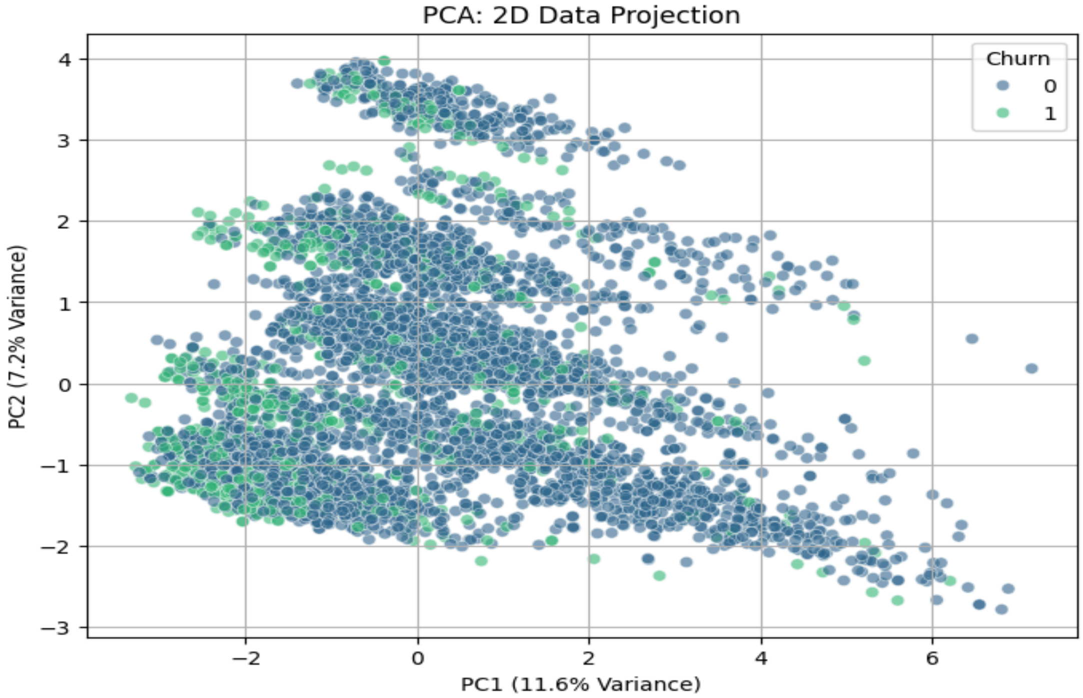
PCA was performed on the scaled training data to analyze the dataset's dimensionality and potential for visualization.

Figure 4: PCA Cumulative Explained Variance



The cumulative variance plot (Figure 4) shows that 22 of the 27 principal components are required to explain 95% of the variance. This indicates that the dataset is high-dimensional and cannot be simplified to just a few components without significant information loss.

Figure 5: PCA 2D Projection



A 2D projection (Figure 5) of the data onto its first two principal components shows a heavy overlap between churned and retained customers. This confirms that the separation boundary is complex and non-linear, justifying the choice of a robust, non-linear model.

5.3 Random Forest

Given the non-linear nature of the data and the class imbalance, a Random Forest classifier was selected. This model is an ensemble of decision trees, which is robust to overfitting and can handle complex interactions.

The model was instantiated with n_estimators=200 and, most importantly, class_weight='balanced'. This setting automatically adjusts model weights to give more importance to the minority class (churners), effectively addressing the data imbalance during the training process.

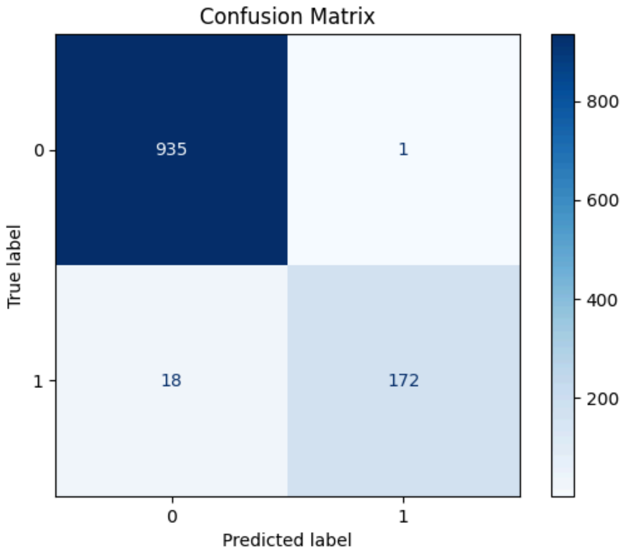
5.4 Model Evaluation

The model was trained on the training set and evaluated on the unseen test set.

Table 2: Random Forest Classification Report

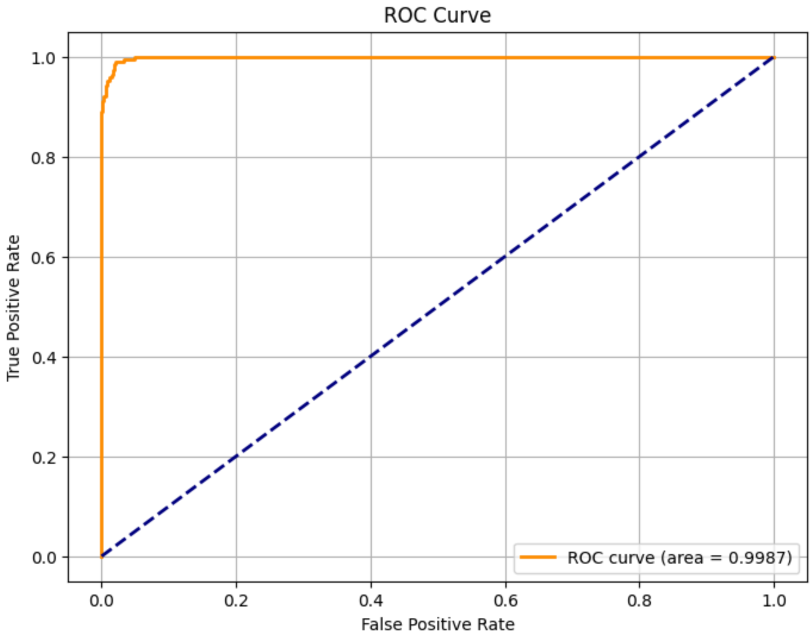
Metric	Class 0 (Churn)	Class 1 (Retain)	Average
Precision	0.98	0.99	-
Recall	1.00	0.91	-
F1-Score	0.99	0.95	-
Accuracy	-	-	0.98

Figure 6: Confusion Matrix



The confusion matrix (Figure 6) provides a clear view of the model's performance. Out of 190 actual churners in the test set, the model correctly identified 172 (True Positives) and missed 18 (False Negatives). This low False Negative rate is crucial for a churn model, as it minimizes the number of at-risk customers who are missed.

Figure 7: Receiver Operating Characteristic (ROC) Curve



The Receiver Operating Characteristic (ROC) curve (Figure 7) plots the True Positive Rate against the False Positive Rate. The curve sits in the top-left corner, indicating high performance. The Area Under the Curve (AUC) is 0.9987, which signifies an outstanding level of discrimination between the two classes.

6. Result and Discussion

EDA revealed that **Tenure**, **Complain**, and **SatisfactionScore** are primary churn drivers; several behavioral segments (e.g., **Mobile Phone** order category, **COD** payment mode, **Computer** login device, **Single** marital status) exhibit elevated risk. The PCA analysis confirmed limited linear separability; the Random Forest capitalized on non-linear patterns to achieve excellent discrimination (ROC-AUC ≈ **0.999**).

Business-wise, the **high recall on churners** means fewer at-risk customers slip through. With calibrated probabilities, stakeholders can set action thresholds to trigger offers, service outreach, or UX fixes.

7. Conclusion and Future Work

This study successfully developed a high-performance Random Forest model capable of predicting customer churn in an e-commerce environment with 98% accuracy and an ROC-AUC of 0.9987. Key drivers identified through EDA, such as Tenure, Complain, and SatisfactionScore, were shown to be critical factors.

The model's high recall for the churn class (92%) makes it an effective tool for business. By deploying this model, the company can move from a reactive to a proactive customer retention strategy.

Future work could involve:

1. Model Deployment: Integrating the model into a live dashboard to provide real-time risk scores.
2. Feature Engineering: Creating more complex features, such as "monetary value" or "purchase frequency" ratios.
3. Alternative Models: Testing other advanced ensemble methods, such as XGBoost or LightGBM, to potentially achieve further marginal gains in performance.

Acknowledgments

This study was developed as part of an applied analytics project on predictive modeling for customer churn in the e-commerce industry. I would like to express my gratitude to the Kaggle community for providing open access to the E-Commerce Customer dataset, which served as the foundation for this research. I also wish to acknowledge the developers and contributors of open-source tools, including Python, Pandas, scikit-learn, Matplotlib, and Seaborn, whose technologies enabled efficient data preprocessing, visualization, and machine learning model development throughout the course of this study.

References

Avi S. (2025) Customer Churn Prediction & Retention
https://www.kaggle.com/code/arsri1/customer-churn-predition-retention
Geeksforgeeks, (2025), Random Forest Algorithm in Machine Learning
https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/
Albers U. (2022) K-means Clustering and Principal Component Analysis in 10 Minutes
https://towardsdatascience.com/k-means-clustering-and-principal-component-analysis-in-10-minutes-2c5b69c36b6b/

Appendix A. Code Snippets

Key functions for metrics calculation (Accuracy, Precision, Recall, F1-Score, AUC).
Model training Random Forest
Principal Component Analysis (PCA) for exploration and structure-diagnosis

Appendix B. Additional Figures and Tables

Statistical Summary of Numerical Features (Table 1, Figure 1, Figure 2)
Training/ Testing performance (Table 2, Figure 6)
ROC-AUC Comparison (Figure 7)
PCA 2D Projection (Figure 5)