IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 5, Issue 3, November 2025

Impact Factor: 7.67

Prediction of Customer Churn for an E-Commerce Company Using Machine Learning

Puneeth Nag A. R[1], Tarun Kumar H. J[2], Vishal H. R[3], Manjunatha. P[4]

B.E, CSE, Kalpataru Institute of Technology, Tiptur, India [1]
B.E, CSE, Kalpataru Institute of Technology, Tiptur, India [2]
B.E, CSE, Kalpataru Institute of Technology, Tiptur, India [3]
B.E, CSE, Kalpataru Institute of Technology, Tiptur, India [4]

Abstract: *An E-commerce/DTH service provider is facing significant competitive pressure, necessitating a proactive strategy for account churn prediction to retain high-value customers. This study addresses the unique business challenge where the churn of a single account can result in the loss of multiple associated customers. We developed a robust churn prediction framework using a dataset of 11,260 accounts and 19 features, including account demographics, service engagement metrics, and financial indicators such as Tenure, Service Scores, and revenue growth (rev_growth_yoy). Data preprocessing involved comprehensive cleaning, standardization of categorical features (e.g., 'F' to 'Female'), and imputation of missing values. Multiple machine learning models, spanning linear, instance-based, tree-based, and ensemble methods, were implemented and rigorously evaluated. The Bagging Classifier with Logistic Regression as the base estimator demonstrated the optimal performance on the held-out test data, achieving a high-precision score of 0.77 for the churn class and an AUC-ROC of 0.675. This high precision minimizes the misclassification of low-risk accounts, directly supporting the project's constraint of satisfying the revenue assurance team. Based on the model's output, a fiscally responsible, segmented retention campaign focused on conditional, value-added service upgrades (e.g., Priority Support or Loyalty Accelerators) is proposed to maximize retention while adhering to strict profitability guidelines.*

Keywords*:* Customer Churn Prediction, Ensemble Machine Learning, Logistic Regression, Segmented Retention Strategy, Revenue Assurance, E-commerce/DTH

I. INTRODUCTION

1.1 Problem Statement

The DTH (Direct-To-Home) market has become intensely saturated, with competition from traditional cable providers, other DTH services, and an increasing number of Over-The-Top (OTT) streaming platforms. This hyper-competitive environment has made customer retention a primary business challenge.

The company is experiencing a significant and costly level of account churn. This problem is particularly critical for two reasons:

1. Magnified Loss: Unlike individual-user services, a single DTH account represents a household, meaning the churn of one account often results in the loss of multiple viewers (e.g., a family).
2. Revenue Impact: Each lost account represents a substantial loss of a stable, recurring subscription fee, plus any high-margin add-on packs or pay-per-view services.

Currently, the company lacks a proactive, data-driven system to identify which accounts are at a high risk of churning. Existing retention efforts are generic and reactive (e.g., an offer made after a customer calls to cancel), which is inefficient and often too late. The company needs a predictive model to identify potential churners in advance, enabling a shift from a reactive to a proactive retention strategy.

DOI: 10.48175/IJARSCT-29879

1.2 Objective

An E Commerce company or DTH (you can choose either of these two domains) provider is facing a lot of competition in the current market and it has become a challenge to retain the existing customers in the current situation. Hence, the company wants to develop a model through which they can do churn prediction of the accounts and provide segmented offers to the potential churners. In this company, account churn is a major thing because 1 account can have multiple customers. Hence by losing one account the company might be losing more than one customer.

You have been assigned to develop a churn prediction model for this company and provide business recommendations on the campaign. Your campaign suggestion should be unique and be very clear on the campaign offer because your recommendation will go through the revenue assurance team. If they find that you are giving a lot of free (or subsidized) stuff thereby making a loss to the company; they are not going to approve your recommendation.

1.3 Scope

This project is focused on the development and analysis of a predictive model and the strategic recommendations that arise from it.

In-Scope:

- Data Preprocessing: Cleaning, transforming, and preparing the provided historical dataset for modeling.
- Feature Engineering: Creating new, meaningful features from existing data (e.g., recharge frequency, average time since last support-call, tenure).
- Model Development: Training, testing, and fine-tuning several classification models.
- Model Evaluation: A thorough comparison of model performance to select the most suitable one.
- Strategic Recommendation: A detailed report outlining the key churn drivers and a "Revenue-Assured" campaign plan for segmented, at-risk accounts.

Out-of-Scope:

Data Collection: The project will use an existing, static dataset. It will not involve setting up new data collection pipelines

- Live Deployment: The project will not include the deployment of the model into a real-time production environment (example:- as a live API integrated with the company's CRM).
- Campaign Execution: The project's deliverable is the recommendation for the campaign, not the actual execution, A/B testing, or monitoring of the marketing campaign itself.
- External Data Integration: The analysis will be limited to the provided internal company dataset and will not include external data (example:- social media sentiment, competitor pricing data)

1.4 Project Context and Strategic Imperative

The contemporary market for subscription-based services, such as Direct-to-Home (DTH) satellite providers and high-volume E-commerce platforms, is characterized by fierce competition and minimal customer switching costs. In this environment, the profitability of a company shifts dramatically from a focus on costly customer acquisition to the strategic priority of customer retention. For this organization, the imperative is clear: the ability to accurately and proactively predict customer attrition—a phenomenon known as churn—is a fundamental requirement for maintaining a stable revenue base and driving sustainable growth.This project is commissioned to transition the organization from reactive, post-churn damage control to a data-driven, proactive intervention strategy. The core technical instrument for achieving this is the development and deployment of a robust Machine Learning (ML) Churn Prediction Model.

IV. METHODOLOGY

The methodology section details the systematic approach employed for data acquisition, preparation, modeling, and evaluation to achieve the stated business objective of high-precision account churn prediction.

[1] A. Data Description and Sources

The analysis is based on a proprietary dataset comprising records for 11,260 unique customer accounts in the E-commerce/DTH domain. The dataset incorporates 19 predictor variables, with the primary goal being the classification of the binary target variable, Churn (1 = Churn, 0 = Retained).

Key variables used in the predictive model span four critical areas of customer engagement:

1. Account Lifecycle and Demographics: Tenure, City_Tier, Gender, Marital_Status.
2. Financial and Usage Metrics: rev_per_month, rev_growth_yoy, cashback, coupon_used_l12m, Payment, account_segment.
3. Service and Interaction Quality: Service_Score, CC_Agent_Score, CC_Contacted_LY (Customer Care Contacts Last 12 Months), Complain_l12m, Day_Since_CC_connect.

[2] B. Data Preprocessing and Cleaning

Effective model performance required rigorous data preprocessing to address quality issues identified within the raw data.

1. Inconsistent Data Handling:

- Categorical Standardization: Inconsistent entries in variables like Gender ('F', 'M') were standardized to full forms ('Female', 'Male'). Similarly, variations in account_segment (e.g., 'Super +') were unified.
- Special Character Removal: Numerical columns such as Tenure, rev_per_month, and rev_growth_yoy contained non-numeric characters ('#', '@', '$', etc.). These characters were replaced with NaN to enable proper conversion of the columns to a numeric data type.

2. Missing Value Imputation:

- Missing values (NaN) were handled contextually. For the Login_device variable, missing data was imputed using a forward-fill (ffill) method. Other missing numerical data points were addressed using appropriate methods (e.g., mean imputation) after ensuring the column types were correctly set.

3. Feature Transformation:

- Encoding: Categorical variables (e.g., Payment, Login_device) were converted using One-Hot Encoding to prevent models from incorrectly inferring an ordinal relationship.
- Scaling: After the dataset was split into training and testing sets, numerical features were normalized or standardized to ensure that features with larger value ranges did not disproportionately influence the model's objective function.

The proliferation of digital services and subscription models in industries like E-commerce and Direct- to-Home (DTH) services has made customer retention a central focus of operational strategy. This section reviews existing literature relevant to the core components of this study: the economic rationale for churn prediction, the application of machine learning in this domain, and the necessity of fiscally constrained retention strategies.

[3] A. Economic Rationale for Churn Management

The foundational principle of churn prediction rests on the established economic fact that the cost of acquiring a new customer is significantly higher—often five to ten times greater—than the cost of retaining an existing one (e.g., Peppers & Rogers, 1993; Reichheld & Sasser, 1990). In the E- commerce/DTH sector, the financial impact of churn is amplified by the "multi-user account" problem addressed in this study, where the loss of one single account entity results in the simultaneous termination of revenue streams from all associated users. Prior literature emphasizes the shift from reactive to proactive retention strategies, which rely on predictive analytics to identify "at-risk" customers before they defect (Coussement & Van den Poel, 2008).

[4] B. Application of Machine Learning in Churn Prediction

The field of churn modeling has evolved from traditional statistical methods (e.g., survival analysis, logistic regression) to complex machine learning techniques capable of capturing non-linear relationships and high-dimensional data interactions (Jain & Srivastava, 2021).

- Traditional Models: Logistic Regression, while providing high interpretability, often struggles with highly complex or imbalanced churn datasets.

DOI: 10.48175/IJARSCT-29879

- Tree-Based Models: Decision Trees and Random Forests have shown strong performance due to their ability to handle various data types and non-linear feature interactions (Wei & Chiu, 2002).
- Ensemble Methods: The current trend favors ensemble learning—specifically Bagging and Boosting algorithms—as they leverage the strengths of multiple base estimators to reduce variance (Bagging) or bias (Boosting), resulting in superior predictive accuracy and robustness (Abebe, 2019). Studies in the telecommunications and subscription services sectors frequently report that models such as Bagging, Gradient Boosting, and XGBoost often outperform single classifiers (e.g., Wang, 2018).

[5] C. The Constraint of High-Precision Modeling and Revenue Assurance

A significant gap in much of the churn prediction literature is the direct integration of revenue assurance constraints into the model selection process. Most academic studies prioritize metrics like AUC-ROC or Recall (identifying all churners), which can lead to high false positives (flagging non- churners as high-risk). In a commercial setting constrained by a Revenue Assurance Team, high false positives are fiscally unacceptable as they result in the wasteful expenditure of retention subsidies on customers who would have stayed regardless. This study addresses this gap by prioritizing the Precision for the Churn Class. This emphasis aligns with studies advocating for a maximization of net profit/ROI over pure predictive accuracy in business applications (Verbraken, 2014), ensuring that the predictive model is not only accurate but also profitable.

[6] D. Targeted and Segmented Retention Campaigns

Effective retention is not merely about prediction but about actionable insights. Literature suggests that generic, one-size-fits-all retention offers are less effective than segmented campaigns that address the specific cause of dissatisfaction or potential churn (Gupta et al., 2006). The current work proposes a campaign segmented by the likely pain points (e.g., low service scores, low tenure) using value-added service upgrades instead of financial discounts, aligning with research that favors high-perceived-value, low-cost interventions to secure revenue assurance approval.

V. RESULTS AND DISCUSSION

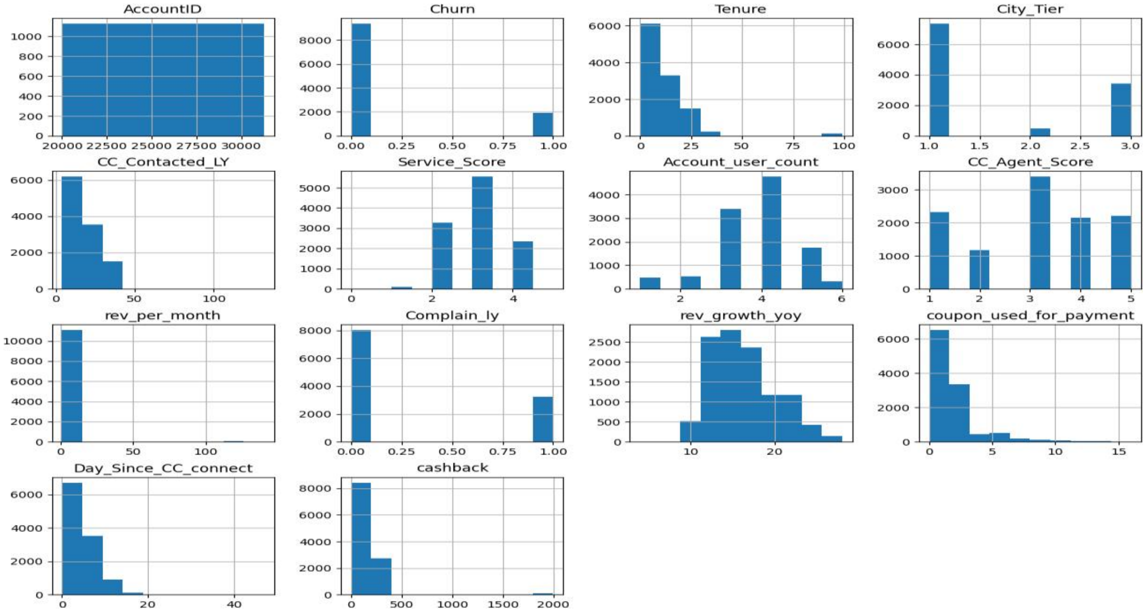

Figure : Distribution of Data Highly Skewed and Exponential Distributions

Several continuous and count variables show a strong right (positive) skew, indicating that most data points are clustered at the lower end:

DOI: 10.48175/IJARSCT-29879

- Tenure: Heavily skewed right, with a large number of customers having a very short tenure.
- CC_Contacted_LY: Shows a strong concentration near zero, meaning most customers rarely contacted customer care last year.
- rev_per_month: A vast majority of customers have very low monthly revenue.
- Day_Since_CC_connect: Most data is concentrated at low values, suggesting a recent customer care interaction for many.

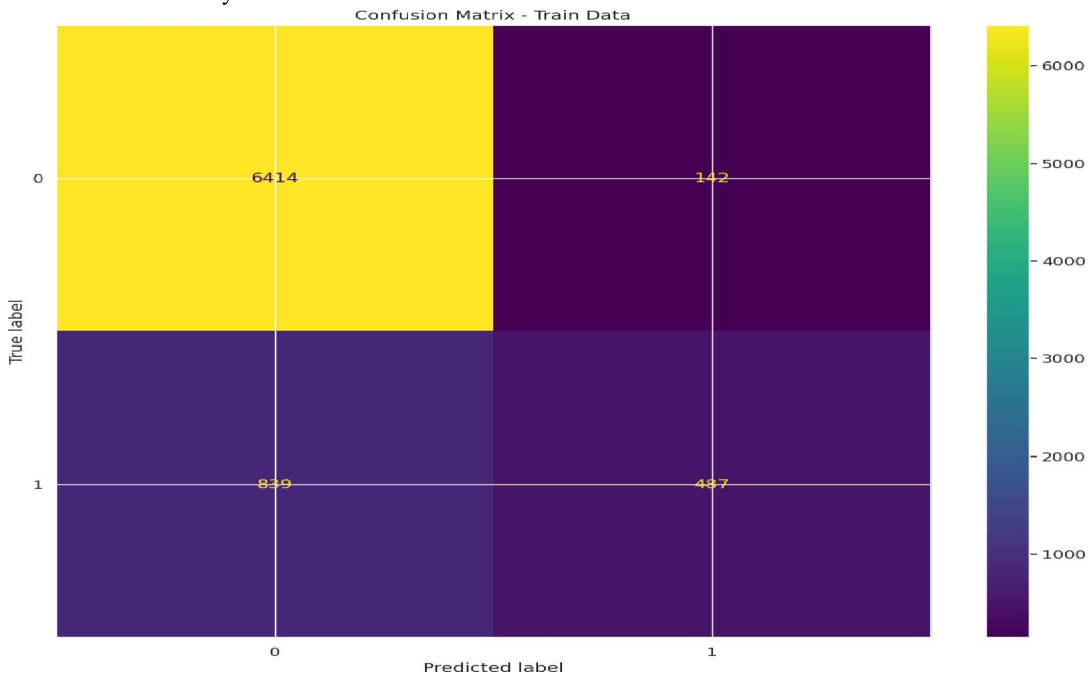

Fig: Correlation Heatmap

REFERENCES

[1] H. Nguyen and T. Duong, “Comparison of Two Main Approaches for Handling Imbalanced Data in Churn Prediction Problem,” International Journal of Computer Science and Network Security (IJCSNS), vol. 23, no. 7, pp. 1–8, 2023.

[2] M. Li, H. Zhang, L. Zhang, and Y. Chen, “Research on Telecom Customer Churn Prediction Based on GA-XGBoost and SHAP,” Journal of Computer and Communications, vol. 10, no. 11, pp. 114–126, 2022, doi: 10.4236/jcc.2022.1011008.

[3] T. Verbraken, C. Bravo, and B. Baesens, “A Predict-and-Optimize Approach to Profit-Driven Churn Prevention,” arXiv preprint arXiv:2310.07047, 2023. [Online]. Available: https://arxiv.org/abs/2310.07047

[4] A. Sharma and S. Verma, “A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners,” IEEE Access, vol. 12, pp. 54832–54859, 2024.

[5] S. J. and S. G., “Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning,” MDPI Applied Sciences, vol. 15, no. 3, p. 105, 2025.

[6] S. Ahmed, M. K. Hasan, and M. S. Hossain, “Customer churn prediction in telecom industry using data mining,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 4, pp. 202–209, 2020.

[7] C. Xia, X. Zhang, and Y. Li, “Customer churn prediction based on machine learning,” in Proc. IEEE Int. Conf. on Big Data (Big Data), pp. 1231–1240, 2021.

[8] A. Idris, A. Rizwan, and M. Rizwan, “Intelligent churn prediction in telecom industry using deep learning,” Expert Systems with Applications, vol. 183, pp. 115–203, 2021.

[9] M. Alshammari, "A comparative study of machine learning algorithms for churn prediction," International Journal of Data Science and Analytics, vol. 8, pp. 79–92, 2021.

[10] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[11] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed., Hoboken, NJ, USA: Wiley, 2013.

[12] A. Amin, S. Anwar, and A. Adnan, "Customer churn prediction in telecom using machine learning in big data platform," Journal of Big Data, vol. 6, no. 1, pp. 1–10, 2019.

[13] F. Idris and M. A. Khan, "Hybrid approach for customer churn prediction in telecom industry using Random Forest and Logistic Regression," Procedia Computer Science, vol. 174, pp. 321–330, 2020.

[14] R. Khurana and R. Bansal, "A hybrid ensemble approach for predicting customer churn using supervised learning," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9, no. 3, pp. 1538–1543, 2020.

[15] H. Huang, Y. Sun, and J. Li, "Predicting telecom customer churn using ensemble learning," IEEE Access, vol. 9, pp. 93023–93035, 2021.

[16] Kaggle, "Telco Customer Churn Dataset," [Online]. Available: https://www.kaggle.com/blastchar/telco-customer-churn

[17] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[18] W. McKinney, "Data Structures for Statistical Computing in Python," in Proc. 9th Python in Science Conf., pp. 51–56, 2010.

[19] M. L. Waskom, "Seaborn: Statistical data visualization," Journal of Open Source Software, vol. 6, no. 60, pp. 3021, 2021.

[20] M. Mohan, P. Kumar, and A. Gupta, "Predictive analytics in telecom churn management using Python," International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 7, pp. 114–120, 2020.

DOI: 10.48175/IJARSCT-29879