Repeater: Publikasi Teknik Informatika dan Jaringan
Volume 3, Number 2, April 2025
E-ISSN : 3046-7276, P-ISSN : 3046-7284; Hal. 17-31
DOI: https://doi.org/10.62951/repeater.v3i2.401
Available online at: https://journal.aptii.or.id/index.php/Repeater

Implementing XGBoost Model for Predicting Customer Churn in E-Commerce Platforms

Andy Hermawan [1]*, Aji Saputra [2], Muhammad Dhika Rafi [3], Syafiq Basmallah [4], Yilmaz Trigumari Syah Putra [5], Wafa Nabila [6]
[1] Universitas Indraprasta PGRI, Indonesia
[2] Universitas Khairun, Indonesia
[3,4,5,6] Purwadhika Digital Technology School, Indonesia

Jl. Jenderal Sudirman No.Kav. 21 10, RT.10/RW.1, Kuningan, Karet, Kecamatan S etiabudi, Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12930
Email : andy.hermawan@unindra.ac.id, aji.saputra@unkhair.ac.id

Abstract*. Customer churn is a major challenge in e-commerce, directly affecting revenue and profit. This study aims to develop a machine learning model using XGBoost to predict churn probability. To handle class imbalance, SMOTE was applied as a resampling method, and hyperparameter tuning was performed to enhance performance. The model was evaluated using the F2-score, prioritizing recall while maintaining precision. The results show that the XGBoost model with SMOTE achieves strong performance, with an F2-score of 0.849 on the tuned test data. This model can help businesses identify at-risk customers early, enabling proactive retention strategies.*

Keywords*: Churn Prediction, E-commerce, Machine Learning, XGBoost.*

1. INTRODUCTION

Customer churn is a measurement of the percentage of customers who stop using a service or canceling a subscription during a certain period, a major challenge in the e-commerce industry as it impacts the business financial and revenue (Li, L., 2022). A high churn percentage suggests that the business is losing too many customers, which could lead to issues like unsatisfactory service, product problems, or poor customer service (Bajaj, S., 2025).

With the rise of machine learning, churn prediction models now provide more accurate and proactive insights into customer behavior. Traditional models like logistic regression and decision trees struggle to capture complex, non-linear customer patterns, while ensemble methods such as XGBoost have emerged as a more advanced solution (Chen et al., 2021). XGBoost efficiently handles large datasets, minimizes overfitting through regularization, and delivers superior predictive performance compared to older techniques. Recent studies confirm that XGBoost consistently outperforms other models in churn classification tasks, making it an ideal choice for businesses looking to implement data-driven retention strategies (Matuszelański, K., 2022).

Beyond just prediction, churn models also drive business decisions by identifying key factors influencing attrition, allowing companies to develop personalized retention campaigns,

Received: February 05, 2025; Revised: February 20, 2025; Accepted: March 10, 2025; Published : March 12, 2025

optimize marketing spend, and enhance customer engagement (Peng et al., 2023). By integrating churn predictions with strategic interventions or programs can effectively reduce customer loss and improve long-term profitability.

Given the growing importance of data-driven retention strategies, this study aims to implement an XGBoost-based churn prediction model for e-commerce platforms, evaluating its performance and providing actionable insights for reducing churn.

2. THEORETICAL STUDY

Customer Churn in E-Commerce

Keeping the customer retention is often a problem faced with many e-commerce platforms. Churn is one of the ways to measure how customers have stopped using one's platform or services. It is important for firms to be aware of customer churn as a high customer churn rate will have a significant impact on business. Costs to acquire new customers could be five times higher than keeping the old one (Dhangar & Anand, 2021).

Machine Learning for Churn Prediction

Machine learning is able to make predictions of the probability of churning. The algorithm of machine learning is trained to recognize the pattern in the dataset in order to make predictions on probability of customer churning. The conservative approaches to churn prediction depend on demographic, product-usage, and revenue features alone (De, 2022). With the help of machine learning, businesses can also understand the features that play big roles in customer churning through feature importances (Fayrix, n.d.). Predictions are made before customers actually churn which will help businesses to mitigate this problem. This can help businesses to decrease customer churn rate and increase customer retention rate which will result in increasing profits.

Application of XGBoost in Churn Prediction

XGBoost is a gradient-boosting-based model that is constructed from many decision trees that are commonly used to address both classification and regression problems. The XGBoost model sequentially trains decision trees using training data. To increase the value of the goal function, the algorithm appends a new decision tree to the earlier decision trees at each iteration (Aydin, 2021). This gradient-boosting framework can improve weak performance from previous base models and provide more robust performance for more accurate predictions (Cai, 2023). Vasudevan et al. (2022) demonstrated that the implementation of XGBoost gives comparatively

more accurate predictions than other learning models. XGBoost is often found to be the best-performing model for imbalanced datasets, as demonstrated by its superior performance on two out of three real datasets (Lai et al., 2021).

3. RESEARCH METHODOLOGY

Table 1. E Commerce Dataset

Number	Features	Data Format	Description
0	Tenure	Numerical	Duration as a customer with the company.
1	WarehouseToHome	Numerical	warehouse to the customer's home distance.
2	NumberOfDeviceRegistered	Numerical	Total registered devices per customer.
3	PreferredOrderCat	Categorical	Preferred category in the last month.
4	SatisfactionScore	Numerical	Satisfaction score regarding the service.
5	MaritalStatus	Categorical	Customer's marital status.
6	NumberOfAddress	Numerical	Total addresses per customer.
7	Complaint	Numerical	Number of complaints in the last 30 days.
8	DaySinceLastOrder	Numerical	Number of days since the last order.
9	CashbackAmount	Numerical	Average cashback amount in the last month.
10	Churn	Numerical	Churn Status label

As seen from the table above, there are various features from the dataset. Each feature has either numerical or categorical data format and has its own function in data analysis. These variables will later be used in predictive modeling to identify key factors contributing to customer churn.

A thing to consider from this dataset is the distribution of churn. From the available dataset, it has 83.7% of customers classified as "Not Churn" and only 16.3% as "Churn," which indicates that the data has a class imbalance. Ignorance of the class imbalance will lead to unfairness and even inadequate generalization abilities (Wu, 2021). The optimal approach is to alter the churned and kept data to achieve a comparable proportion (Ahn, 2020). To handle this issue, later on the use of resampling will be used to improve the model.

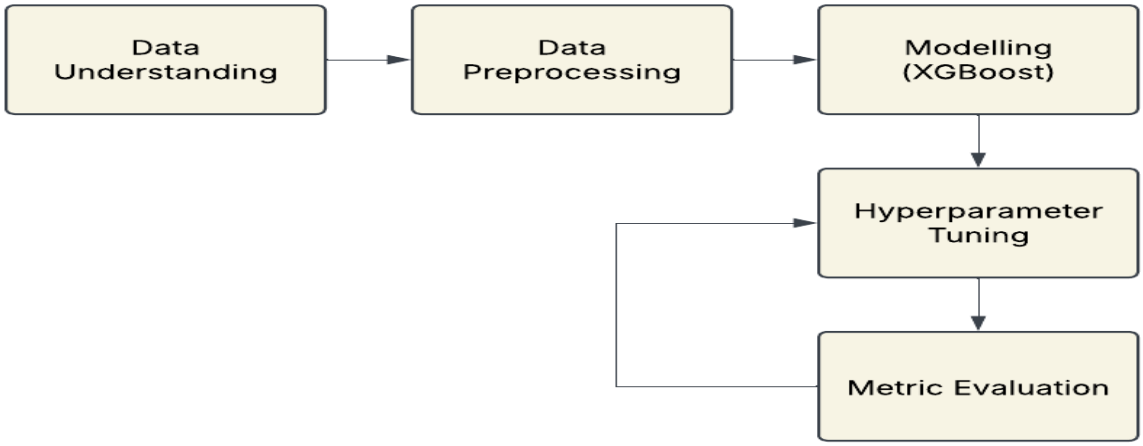
Before applying anything, the data is split into train data and validation data (80% - 20% validation), this is done to prevent data leakage (Nassar, 2023). Data Preprocessing is performed to clean and transform the train dataset into a suitable format for machine learning and to retain the quality of the data (Hermawan et al., 2024). This includes handling missing values, encoding categorical variables, scaling numerical features, and handling imbalance.

The model is resampled using Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance, ensuring that the XGBoost algorithm learns effectively without biasing predictions toward the majority class.This technique generates synthetic samples for the minority class, improving the model's ability to detect customer churn accurately (Chawla et al., 2002). The combination of SMOTE and XGBoost has been shown to enhance predictive performance, particularly in datasets where the number of churners is significantly lower than non-churners (Imani, 2025). Hyperparameter tuning is done using GridSearchCV to find the best parameter in the XGBoost model for optimal metric results.

Overall, the careful preprocessing of data, strategic handling of class imbalance using SMOTE method, selection of the XGBoost algorithm, and thorough hyperparameter optimization constitute a comprehensive approach for building an effective customer churn prediction model. Data cleaning and encoding ensure meaningful input features, imbalance correction addresses the skewed nature of churn events, XGBoost provides a powerful learning framework, and hyperparameter tuning refines the model to find the best metrics (Gan., 2022). The resulting model is not only accurate but also generalizes well, enabling e-commerce businesses to proactively identify and retain at-risk customers with greater confidence.

Research Design

The flowchart below is provided to illustrate the analytical process and enhance understanding of the methodology used in this study.

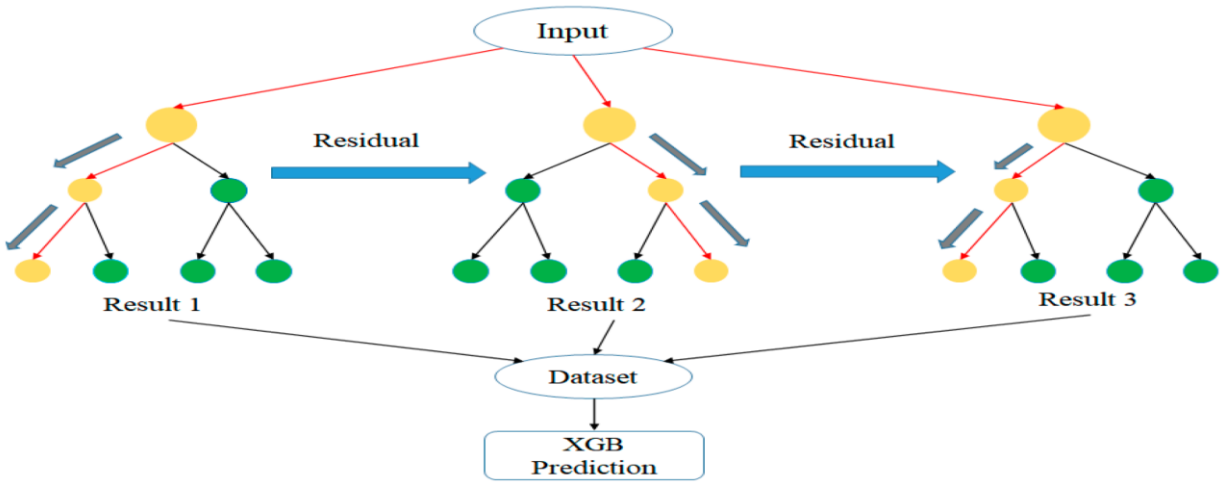


Picture 1. Methodology Flowchart

The flowchart represents a structured approach to churn prediction using a modified CRISP methodology. It starts with Data Understanding, followed by Data Preprocessing to prepare the data for modeling. The Modeling phase focuses on training an XGBoost model, which then undergoes Hyperparameter Tuning to optimize performance. Finally, Metric Evaluation assesses the model's effectiveness, with an iterative loop for further tuning if needed.

Theoretical Foundation

XGBoost



Picture 2. XGBoost Algorithm

Extreme Gradient Boosting (XGBoost) is a machine learning algorithm that's widely used for classification and regression tasks. It builds multiple decision trees in a way that optimizes performance and accuracy while minimizing errors. XGBoost works by combining multiple weak models (decision trees) into a strong predictive model. The model uses gradient boosting to improve predictions iteratively, by adjusting for errors in each step (Bentéjac, 2019). Considering that XGBoost is focused only on decision trees as base classifiers, a variation of the loss function is used to control the complexity of the trees

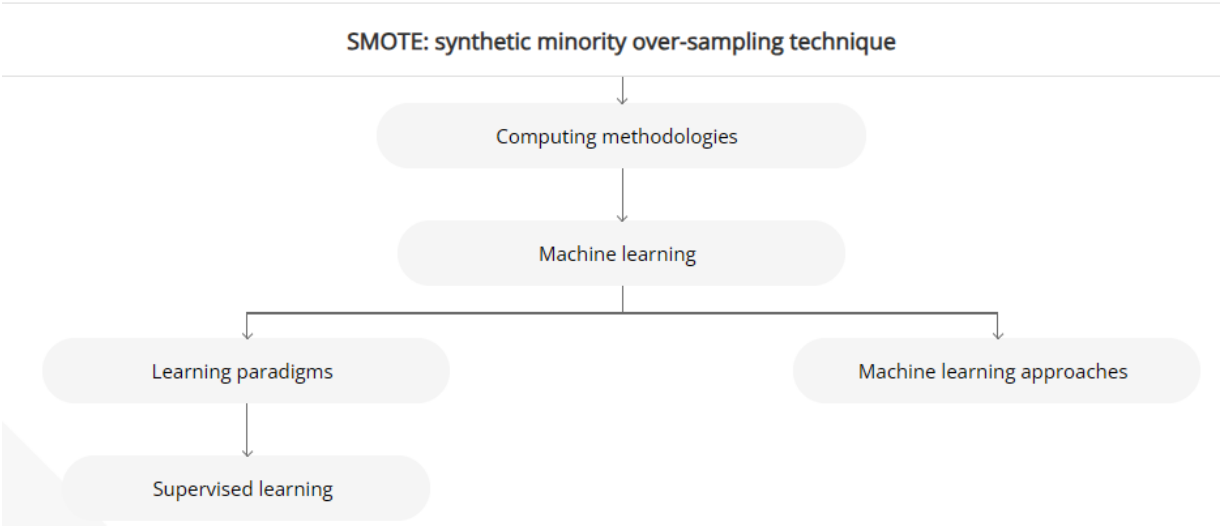
$$L_{xgb} = \sum_{i=1}^{N} L(y_i, F(\mathbf{x}_i)) + \sum_{m=1}^{M} \Omega(h_m)$$

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda \|w\|^2 ,$$

Picture 3. Objective Function of XGBoost

where T is the number of leaves of the tree and w are the output scores of the leaves. This loss function can be integrated into the split criterion of decision trees leading to a pre-pruning strategy. Higher values of γ result in simpler trees. The value of γ controls the minimum loss reduction gain needed to split an internal node (Tang, 2016; Demsar, 2006).

SMOTE Resampling



Picture 3. SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) is a method to address class imbalance in datasets, particularly in binary classification problems where the minority class is underrepresented. Introduced by Chawla et al. (2002), SMOTE enhances the minority class by generating synthetic samples, thereby balancing the class distribution and improving the performance of machine learning models. Mathematically, the generation of a synthetic sample can be expressed as:

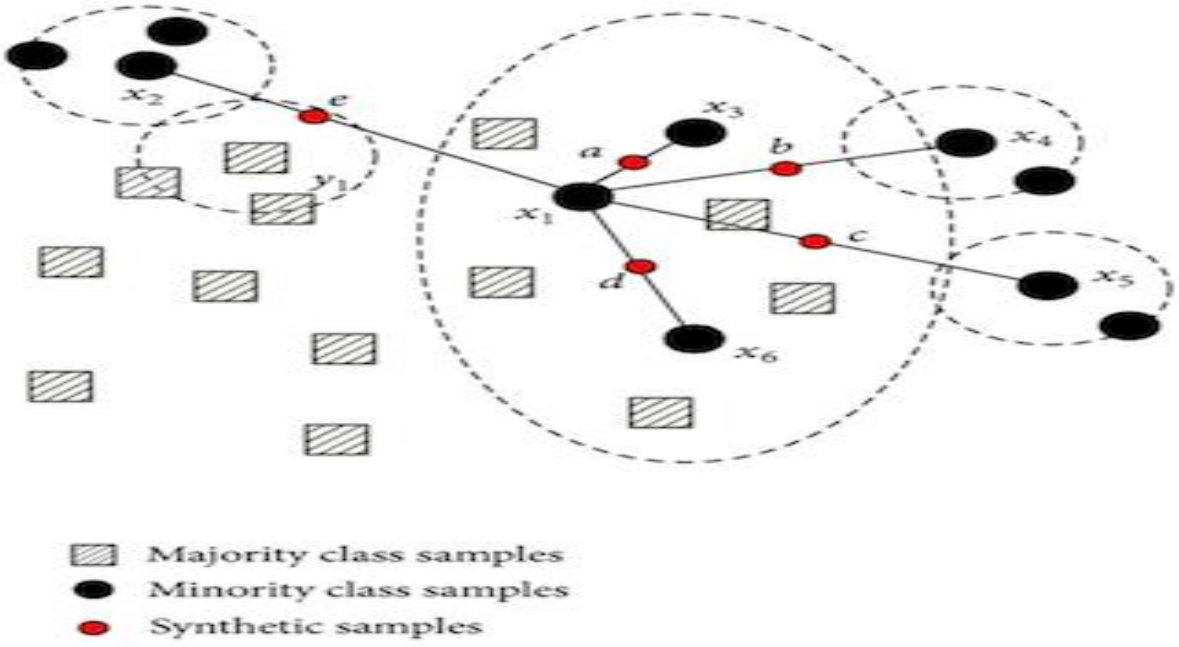
$$x'_i = x_i + \lambda(x_j - x_i)$$

Picture 4. Linear Interpolation Formula

How SMOTE Works:

- SMOTE begins by selecting instances from the minority class in the dataset.
- For each selected minority instance, the algorithm identifies its k nearest neighbors (KNN) within the feature space.

- Synthetic instances are created by interpolating between the selected instance and its neighbors. Specifically, a random neighbor is chosen, and a new instance is generated along the line segment connecting the two, at a randomly chosen point.



Picture 5. SMOTE Workflow

4. RESULTS AND DISCUSSION

Optimal Hyperparameters

Through extensive iterative experimentation, the optimal hyperparameters have been identified, as detailed below.

Table 2. Hyperparameters

Hyperparameter	Best Value	Interpretation
colsample_bytree	0.45	Only **45% of features** are randomly selected per tree, reducing overfitting and improving generalization.
gamma	0.1	A **minimum loss reduction of 0.1** is required to make a split, preventing unnecessary splits and reducing overfitting.
learning_rate	0.035	A **small step size** for updating weights, leading to slower but more stable convergence and reducing the risk of overfitting.
max_depth	5	Trees can grow up to **5 levels**, balancing complexity and generalization to avoid overfitting.
n_estimators	300	The model builds **300 trees**, ensuring sufficient learning while preventing excessive complexity.
reg_alpha (L1 regularization)	0.6	Encourages **sparsity in tree splits**, helping to reduce overfitting by penalizing large coefficients.
reg_lambda (L2 regularization)	0.9	Controls **shrinkage of feature weights**, reducing overfitting by making the model more robust.
scale_pos_weight	3	Balances class weights by **giving 3x more importance** to the minority (churned) class, improving recall for churn detection.
subsample	0.9	Each tree is trained on **90% of the data**, reducing variance while maintaining good performance.

Classification Report of Dataset

Train Data Classification Report:

Table 3. Classification Report

	precision	recall	f1-score	support
0	0.99	0.94	0.96	2189
1	0.75	0.94	0.83	427
accuracy			0.94	2616
macro avg	0.87	0.94	0.90	2616
Weighted avg	0.95	0.94	0.94	2616

F2 Score: 0.8917

Test Data Classification Report:

	precision	recall	f1-score	support
0	0.98	0.91	0.94	547
1	0.66	0.92	0.77	107
accuracy			0.91	654
macro avg	0.82	0.91	0.85	654
Weighted avg	0.93	0.91	0.91	654

F2 Score: 0.8492

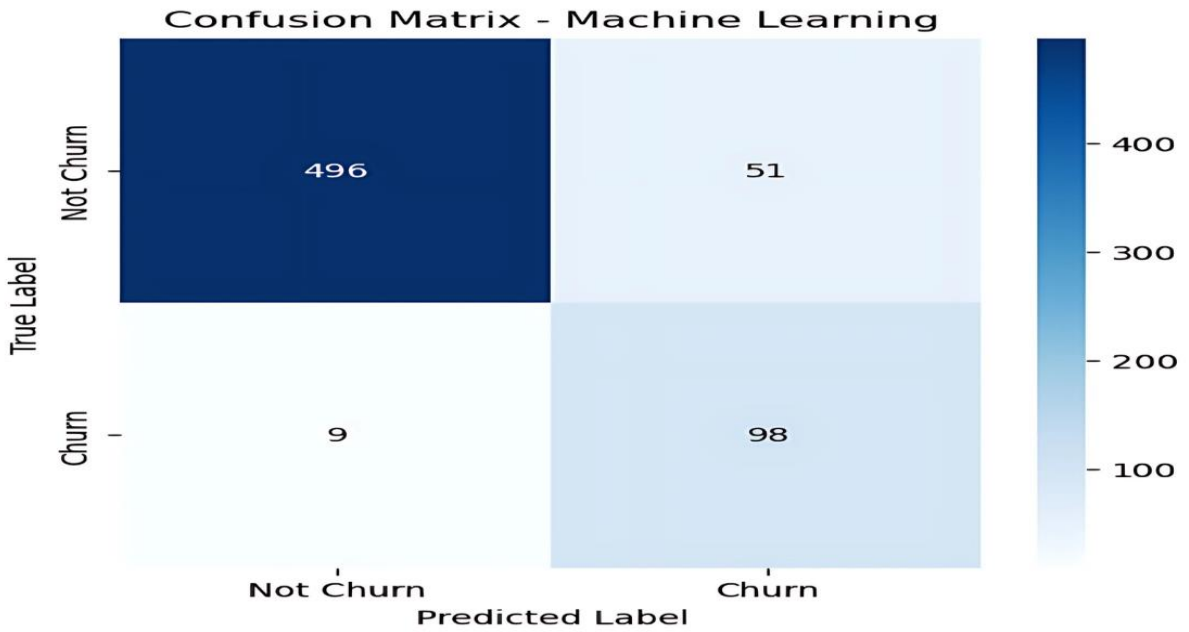
The classification report provides the model's performance results on the train and test datasets, showing its predictive ability. On the train data, the model achieved 94% accuracy, with the majority class (non-churn, label 0) having a precision of 0.99 and a recall of 0.94, resulting in an F1 score of 0.96. This means that the model is very effective in correctly capturing non-churn customers. However, the precision for churned customers is only 0.75, indicating a tendency for false positives, where some non-churn customers are incorrectly classified as churned customers. The macro average F1 score is 0.90, indicating a balanced performance across both classes, while the weighted average F1 score is 0.94, indicating the overall effectiveness of the model. The F2

score of 0.8917 on the train set further strengthens the model's performance on recall, ensuring that churned customers are effectively captured.

On the test data, the model maintains a 91% accuracy, demonstrating good generalization with minimal overfitting. The precision for non-churn customers remains high at 0.98, while recall is slightly lower at 0.91, resulting in an F1-score of 0.94. For churners, recall remains high at 0.92, confirming that the model effectively captures most churn cases. However, precision drops to 0.66, meaning that the model still misclassifies a significant portion of non-churn customers as churners. The macro average F1-score of 0.85 suggests a slight imbalance in prediction performance across classes, while the weighted average F1-score of 0.91 confirms strong overall classification ability. The F2-score of 0.8492 on the test set further validates the model's prioritization of recall, ensuring that actual churners are detected at a high rate.

These results indicate that while the model performs well in identifying churners, improving precision for the minority class could further enhance its reliability. Given that F2-score was used as the primary evaluation metric, the model was optimized to favor recall, which is essential in churn prediction. However, further refinements, such as adjusting the decision threshold or exploring additional predictive features, may help reduce false positives while maintaining strong recall performance.

Confusion Matrix



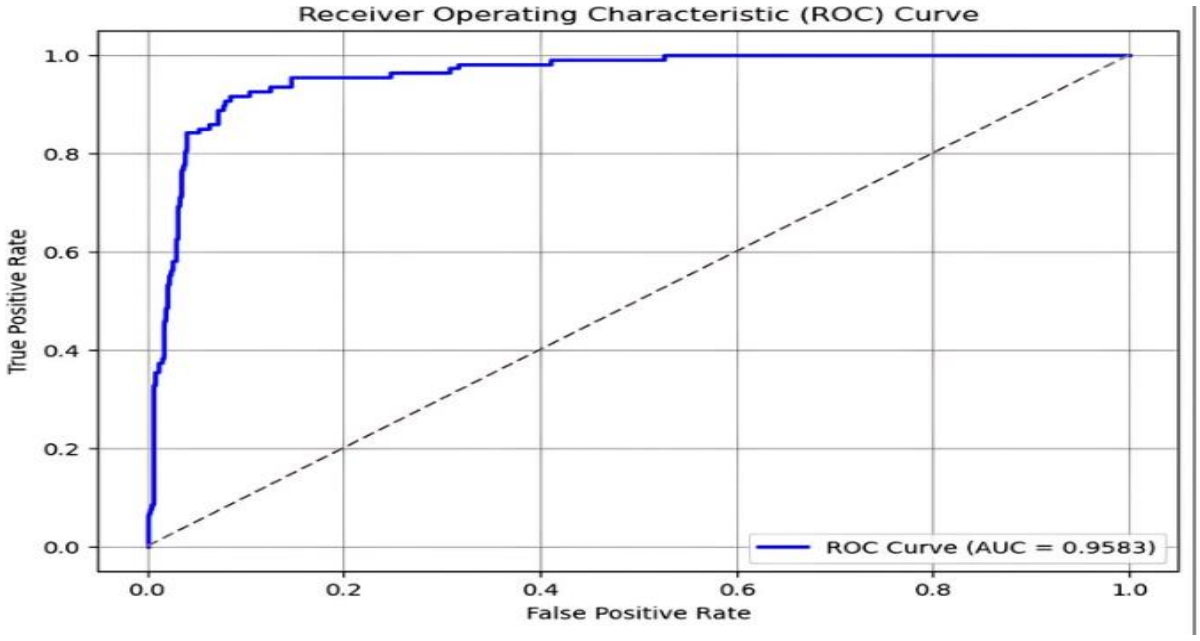
Picture 7. Confusion Matrix

The confusion matrix indicates that the model demonstrates high specificity, as evidenced by the 496 true negatives, meaning that the model is proficient in accurately identifying customers who are not at risk of churn. Additionally, the model's ability to correctly classify 98 churners translates into a high recall, which is critical in churn prediction scenarios where minimizing false negatives is often prioritized to avoid the loss of potentially valuable customers.

However, the presence of 51 false positives suggests a trade-off between recall and precision. While the model effectively captures churners, the moderate precision indicates that some non-churn customers are incorrectly flagged as churners. This could lead to inefficiencies in customer retention strategies, as resources might be allocated to retain customers who do not require intervention.

The low false negative rate (9 cases) is a positive outcome, highlighting the model's capability to minimize missed churn predictions. This aspect is particularly valuable in business contexts where identifying potential churners early enables targeted retention efforts, thereby supporting customer lifetime value (CLV) and reducing churn-related losses.

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve)

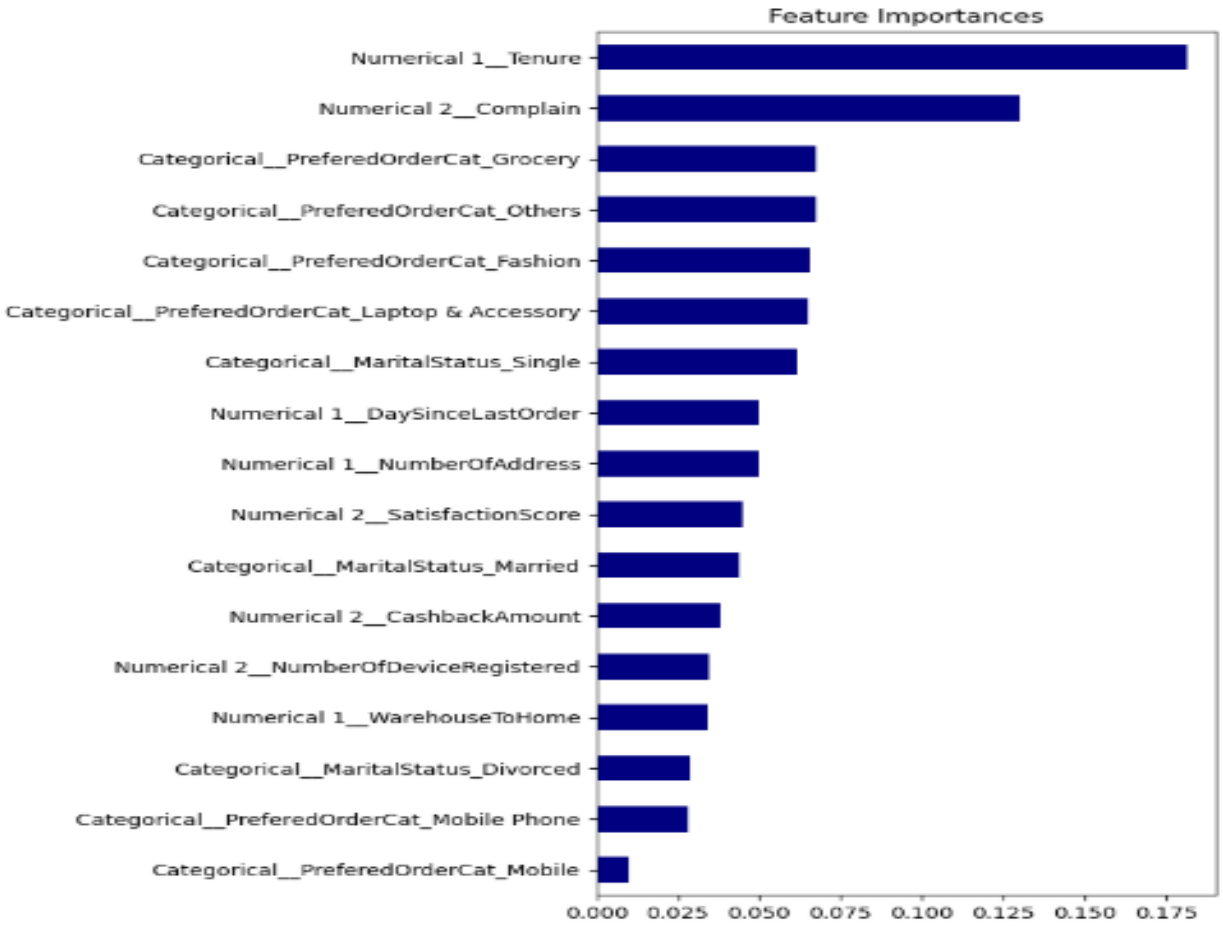


Picture 8. ROC AUC Curve

The Receiver Operating Characteristic (ROC) curve demonstrates the strong performance of the model, with an Area Under the Curve (AUC) of 0.9583. This high AUC value indicates excellent discriminatory ability, suggesting that the model effectively distinguishes between

positive and negative classes. The curve's proximity to the top-left corner of the graph reflects a high true positive rate and a low false positive rate, highlighting the model's robustness and suitability for the classification task.

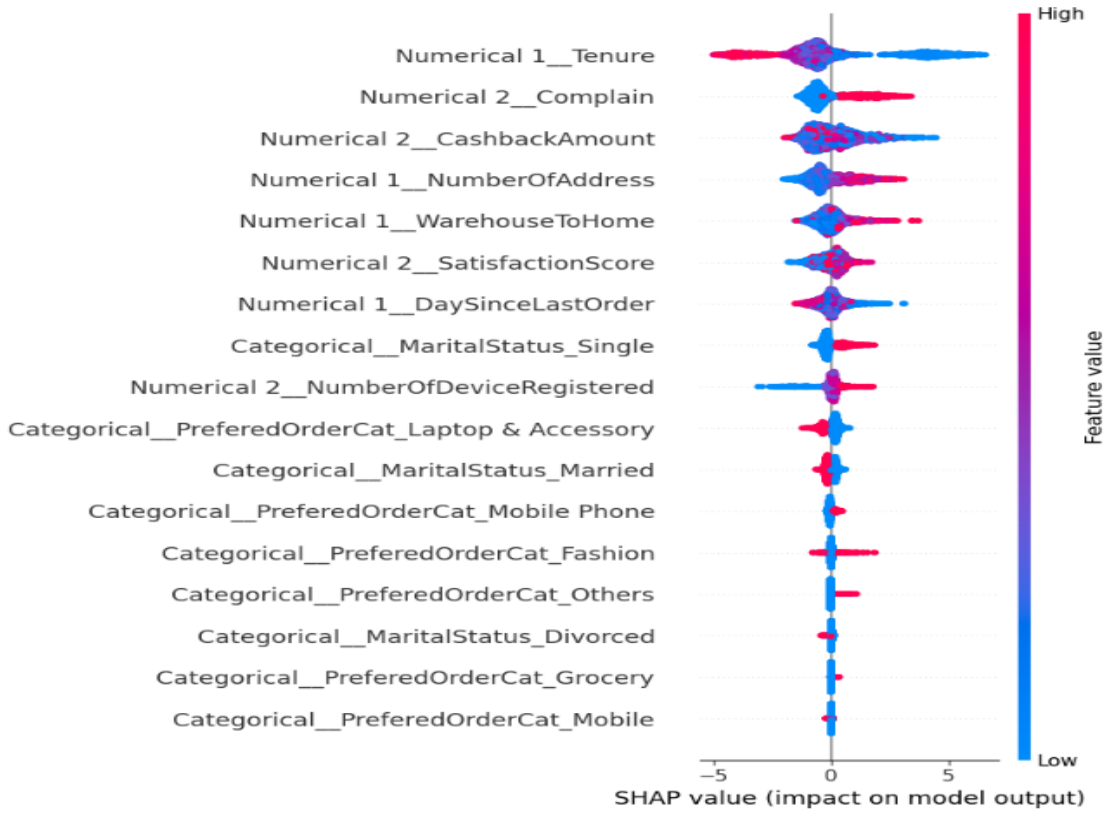
Feature Importance



Picture 9. Feature Importance

The feature importance chart reveals that "Numerical 1__Tenure" and "Numerical 2__Complain" are the most influential features in the model, significantly contributing to prediction accuracy. Categorical variables related to preferred order categories, such as "Grocery," "Others," and "Fashion," also play important roles. The mix of numerical and categorical features among the top contributors indicates that both historical numerical data and categorical preferences are crucial for the model's decision-making process.

SHAP Value



Picture 10. SHAP Value

The SHAP value plot illustrates that "Numerical 1__Tenure" and "Numerical 2__Complain" are the most impactful features on the model's predictions, with higher feature values generally pushing the prediction in a specific direction. The spread of colors indicates how both low and high feature values influence the model output, highlighting the nuanced effect of categorical and numerical features, such as "Numerical 2__CashbackAmount" and "Numerical 1__NumberOfAddress," on classification decisions.

5. CONCLUSION AND RECOMMENDATION

Overall, the findings indicate that the XGBoost-based model is highly effective in predicting customer churn, offering valuable insights for proactive retention strategies. While the model performs well in recall and overall classification, further improvements in precision could enhance efficiency in resource allocation for customer retention. Future work may explore techniques such as threshold tuning, cost-sensitive learning, and additional feature engineering to refine the balance between recall and precision. These improvements will further optimize the model’s practical applicability in e-commerce churn prediction and customer retention strategies.

The XGBoost model implemented in this investigation has the capacity to improve the prediction of consumers who are at risk of churn in comparison to conventional rule-based methodologies. In order to proactively mitigate attrition risks, companies should incorporate predictive analytics into their customer relationship management (CRM) systems. Businesses can optimize their marketing and retention budgets by focusing on high-risk customers, thereby reducing extraneous promotional expenses and maximizing customer retention efforts. The analysis results suggest that the most significant factors in the prediction of attrition are Tenure and Complain.

The current model shows a precision of 66%, indicating a tendency to incorrectly predict non-churn customers as churners. To improve precision, focus on balancing the True Positive rate while minimizing the False Negative and False Positive outcomes. Experimenting with different classification thresholds and model tuning techniques can help achieve this balance.

REFERENCES

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A survey on churn analysis in various business domains. *IEEE Access, 8*, 220816–220839. https://doi.org/10.1109/access.2020.3042657

Aydin, Z. E., & Ozturk, Z. K. (2021). Performance analysis of XGBoost classifier with missing data. *Manchester Journal of Artificial Intelligence and Applied Sciences (MJAIAS, 2*(2). ICMI 2021.

Bajaj, S. (2025). Churn rate 101: Meaning, calculation & reduction strategies. *Shiprocket.* https://www.shiprocket.in/blog/churn-rate/

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). A comparative analysis of XGBoost. *arXiv.org.* https://arxiv.org/abs/1911.01914

Cai, J. (2024). The causes of bank customer churn based on XGBoost and LightGBM models: The evidence from the Kaggle Dataset. *Finance & Economics, 1*(4). https://doi.org/10.61173/ggw4ga77

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, H., Tang, Q., Wei, Y., & Song, M. (2021). Churn prediction model of telecom users based on XGBoost. *Journal on Artificial Intelligence, 3*(3), 115–121. https://doi.org/10.32604/jai.2021.026851

De, S., & Prabu, P. (2022). Predicting customer churn: A systematic literature review. *Journal of Discrete Mathematical Sciences and Cryptography, 25*(7), 1965–1985. https://doi.org/10.1080/09720529.2022.2133238

Dhangar, K., & Anand, P. (2021). A review on customer churn prediction using machine learning approach. *Novateur Publications International Journal of Innovations in Engineering Research and Technology, 8*(5), 193–201.

Fayrix. (n.d.). How churn prediction using machine learning benefits different industries. *Fayrix.* https://fayrix.com/blog/customer-churn-prediction-benefits#content

Gan, L. (2022). XGBoost-based e-commerce customer loss prediction. *Computational Intelligence and Neuroscience, 2022,* 1–10. https://doi.org/10.1155/2022/1858300

Hermawan, A., Jayanti, N. R., Tabaruk, Z., Triadi, F. L. Y., Saputra, A., & Syachrudin, M. R. H. (2024). Membangun model prediksi churn pelanggan yang akurat. *Merkurius: Jurnal Riset Sistem Informasi Dan Teknik Informatika, 2*(6), 67–81. https://doi.org/10.61132/merkurius.v2i6.398

Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS upsampling under varying imbalance levels. *Preprints*. https://doi.org/10.20944/preprints202501.2274.v1

Janez, D. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7,* 1–30.

Lai, S. B. S., Shahri, N. H. N. B. M., Mohamad, M. B., Rahman, H. A. B. A., & Rambli, A. B. (2021). Comparing the performance of AdaBoost, XGBoost, and logistic regression for imbalanced data. *Mathematics and Statistics, 9*(3), 379–385.

Lemmens, A., & Gupta, S. (2017). Managing churn to maximize profits. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.2964906

Li, L. (2022). Research on improved XGBoost algorithm for big data analysis of e-commerce customer churn. *International Journal of Advanced Computer Science and Applications, 13*(12), 1086–1094. https://doi.org/10.14569/IJACSA.2022.01312124

Matuszelański, K., & Kopczewska, K. (2022). Customer churn in retail e-commerce business: Spatial and machine learning approach. *Journal of Theoretical and Applied Electronic Commerce Research, 17*(1), 165–198. https://doi.org/10.3390/jtaer17010009

Nassar, O. (2023). Data leakage in machine learning. *ResearchGate.* https://doi.org/10.13140/RG.2.2.27468.59528

Peng, K., Peng, Y., & Li, W. (2023). Research on customer churn prediction and model interpretability analysis. *PLoS ONE, 18*(12), e0289724. https://doi.org/10.1371/journal.pone.0289724

Pondel, M., Wuczyński, M., Gryncewicz, W., Łysik, Ł., Hernes, M., Rot, A., & Kozina, A. (2021). Deep learning for customer churn prediction in e-commerce decision support. *Business Information Systems*, 3–12. https://doi.org/10.52825/bis.v1i.42

Tang, Q., Xia, G., Zhang, X., & Long, F. (2020). A customer churn prediction model based on XGBoost and MLP. *2020 International Conference on Computer Engineering and Application (ICCEA)*, 608–612. https://doi.org/10.1109/iccea50009.2020.00133

Vasudevan, M., Narayanan, R. S., Nakeeb, S. F., & Abhishek, A. (2022). Customer churn analysis using XGBoosted decision trees. *Indonesian Journal of Electrical Engineering and Computer Science*, *25*(1), 488. https://doi.org/10.11591/ijeecs.v25.i1.pp488-495

Wu, O. (2023). Rethinking class imbalance in machine learning. *arXiv preprint, arXiv:2305.03900.* https://doi.org/10.48550/arXiv.2305.03900