KNNensemble: Custom Oversampling Meets Anomaly-Aware Fusion for Churn Detection

Thuan Bui
Swinburne, FPT University
Hanoi, Vietnam
buiminhthuan1314@gmail.com

Anh Binh Le
International School, VNU
Hanoi, Vietnam
anhbinhle@vnuis.edu.vn

Dung Nguyen
Swinburne, FPT University
Hanoi, Vietnam
nguyendinhdung7a2@gmail.com

An Nguyen
Swinburne, FPT University
Hanoi, Vietnam
chesieanthien@gmail.com

***Abstract*—Customer churn prediction is a critical task for businesses seeking to retain valuable users, yet remains hindered by extreme class imbalance where churned customers make up only a small minority. Traditional supervised models often fail to identify these rare cases, and common oversampling methods such as ADASYN or SMOTE can introduce synthetic noise or fail to prioritize harder-to-learn regions. To address this, we propose KNNensemble, a hybrid learning approach that integrates custom KNN-based oversampling with a weighted ensemble of Random Forest and Isolation Forest. Our method generates minority-class samples by interpolating between real churn cases and their local neighbors, ensuring controlled sample diversity and density. Predictions are then refined through a probabilistic fusion of supervised and unsupervised outputs, with the ensemble weight tuned to optimize F1-score. Experiments on a real-world churn dataset show that KNNensemble achieves a notable improvement in detecting churned users, reaching 10 times improvement in F1-score compared to the baseline model on the minority class — outperforming both baseline classifiers and standard resampling techniques.**

***Index Terms*—Customer Churn Prediction, Imbalanced Classification, Oversampling Techniques, Weighted Fusion, Synthetic Minority Generation.**

I. Introduction

Customer churn prediction has become a key area of research in predictive analytics, particularly for industries with subscription-based business models such as telecommunications, banking, and e-commerce. Churned customers often represent a small but financially significant subset, and effectively predicting their departure enables companies to take proactive retention measures. Studies have shown that churn can account for a considerable proportion of customer lifetime value loss, motivating the development of models that can accurately detect churn under real-world constraints [1].

In the banking and finance sector, the management and prediction of customer churn represents a critical strategic challenge, directly impacting operational efficiency and the organization's growth trajectory. The economic ramifications of churn within the banking sector are significant; notably, the cost of acquiring a new customer is often substantially higher than the cost of retaining an existing one, positioning the reduction of churn rates as a top priority [2]. The ability to accurately predict churn risk enables banks to proactively implement timely intervention measures, thus minimizing revenue leakage and preserving Customer Lifetime Value. A specific characteristic of the banking sector is that the departure of a customer is often not simply the discontinuation of using a single product; it typically entails the loss of the entire relationship across a variety of services (such as checking accounts, savings, loans, credit cards, investments, etc.), significantly exacerbating financial losses.

Most existing churn prediction systems rely on supervised learning methods trained on historical labeled data. Algorithms such as decision trees, logistic regression, support vector machines, and ensemble models like Random Forest are widely used [3]. However, a recurring challenge in churn datasets is the severe imbalance between churned and non-churned customers. This imbalance tends to bias classifiers toward the majority class, reducing recall for the minority churn class. To mitigate this, oversampling techniques such as SMOTE [4] and ADASYN [5] have been proposed to synthetically generate minority class examples. In parallel, unsupervised approaches like Isolation Forest [6] have gained attention for their ability to detect anomalous behavior without the need for labeled data.

Nevertheless, these approaches face notable limitations. Oversampling methods often lack fine control over sample placement and may generate noisy or unrealistic synthetic data, especially near decision boundaries. Supervised classifiers, even with class balancing, struggle to generalize well in the presence of extreme skew. Meanwhile, unsupervised models typically exhibit low precision and limited interpretability when deployed independently. These challenges highlight the need for a hybrid modeling strategy that effectively integrates synthetic sample generation with robust ensemble prediction.

II. Related Work

A. Customer Churn Prediction in Banking

Customer churn has been widely investigated in various industries, but its manifestation in the banking sector is notably more complex and gradual compared to churn in telecommunications or e-commerce. In banking, churn behavior often involves progressive disengagement, such as reduction in transaction frequency, inactivity across product lines, or switching to competing institutions, making early identification more difficult.

Several studies have addressed this domain using classical supervised learning. Tran et al. [7] evaluated a wide range of classifiers on banking churn data, identifying decision trees

and ensemble methods as effective due to their interpretability and robustness. Lariviere and Van den Poel [8] introduced a long-term churn prediction model for retail banks using survival analysis and logistic regression, highlighting the importance of temporal signals. More recent work by Tsai and Lu [9] applied neural networks to extract non-linear churn indicators from customer account features.

Yet, these approaches tend to rely heavily on static supervised learning and overlook the dynamic nature of customer relationships. Moreover, few explicitly address the class imbalance inherent in real-world banking churn datasets, which often results in poor recall for the churn class — a critical failure in operational contexts.

B. Handling Class Imbalance in Churn Modeling

Addressing class imbalance has been a major focus in churn research. While techniques such as SMOTE [4] and ADASYN [5] remain popular, they are not without shortcomings. SMOTE interpolates minority samples uniformly without considering local sample distribution, which can create ambiguous class boundaries. ADASYN focuses on difficult examples near the boundary but may amplify noise in overlapping regions.

In response, researchers have explored ensemble and hybrid approaches. Shumaly et al. [10] proposed a boosting-based imbalance learning method that dynamically adjusts sampling weights, achieving better minority class performance in churn datasets. Dong et al. [11] introduced Random-SMOTE, which first clusters minority instances before sampling to preserve local structures. These methods improve upon naive oversampling but require sensitive tuning and often lack generalizability across domains.

Unsupervised learning has also been adopted. Isolation Forest [6] and Local Outlier Factor (LOF) [12] have been used to model churn as a deviation from normal behavior. While promising, such models are often limited by high false positive rates and a lack of semantic interpretability. Recent studies such as Chen et al. [13] have investigated hybrid architectures combining supervised classifiers with anomaly detection outputs, showing that strategic fusion can significantly improve detection of rare events like churn.

Despite these advances, few methods offer precise control over synthetic sample generation or allow adaptive combination of supervised and unsupervised predictions — a gap we aim to address with our proposed framework.

III. Methodology

The proposed framework, **KNNensemble**, addresses the challenge of highly imbalanced churn classification by combining three components: (i) custom minority oversampling guided by local neighbor interpolation, (ii) supervised learning via Random Forest, and (iii) anomaly-aware scoring through Isolation Forest. The final prediction is obtained through a weighted probabilistic ensemble that balances supervised and unsupervised perspectives. Each component is formally described below.

A. KNN-Based Oversampling for Minority Class Augmentation

Let $X = \{x_1, x_2, \ldots, x_n\}$ be the training instances and $y \in \{0, 1\}$ the corresponding churn labels, where $y = 1$ indicates a churned customer. Due to the natural imbalance ($|y = 1| \ll |y = 0|$), directly training classifiers on the raw data leads to strong bias toward the majority class. To address this, a controlled oversampling mechanism is employed to synthetically expand the minority class distribution.

For each minority-class instance $x_i \in X_{\text{min}}$, its k-nearest neighbors within the same class are identified using Euclidean distance in the feature space. A set of synthetic samples is then generated using the interpolation rule:

$$x_{\text{new}} = x_i + \lambda(x_j - x_i), \quad \lambda \sim \mathcal{U}(0, 1), \quad x_j \in \text{KNN}(x_i)$$

This process is repeated iteratively until the minority class reaches a predefined ratio r relative to the majority class. By constraining interpolation within local clusters, this strategy reduces the likelihood of crossing class boundaries or generating out-of-distribution samples.

Unlike ADASYN, which targets borderline samples, the proposed strategy ensures uniform spatial coverage and preserves the internal geometry of the minority class manifold. This localized, data-aware augmentation enhances generalization without relying on external balancing heuristics.

B. Supervised Classification via Random Forest

Following minority augmentation, a Random Forest classifier is trained on the enriched dataset. As an ensemble of decision trees, Random Forest is well-suited for capturing non-linear relationships and offers robustness to overfitting, especially when trained on heterogeneous features.

Each tree in the ensemble performs recursive binary partitioning and outputs a local class vote. The final prediction score for a sample $x \in X_{\text{test}}$ is given by the averaged posterior probability:

$$p_{\text{RF}}(x) = \frac{1}{T} \sum_{t=1}^{T} h_t(x), \quad h_t(x) \in [0, 1]$$

where h_t denotes the prediction from tree t, and T is the total number of trees. This score reflects the likelihood of churn under supervised historical learning and forms the first component of the decision ensemble.

C. Anomaly Detection with Isolation Forest

To introduce a complementary, label-agnostic perspective, an Isolation Forest is trained solely on the feature vectors of the training data. Unlike traditional density-based anomaly detectors, Isolation Forest operates by recursively partitioning the data space through random splits, isolating anomalous observations in fewer steps.

For each test instance x, an anomaly score $a(x) \in \mathbb{R}$ is produced based on the average path length across all isolation

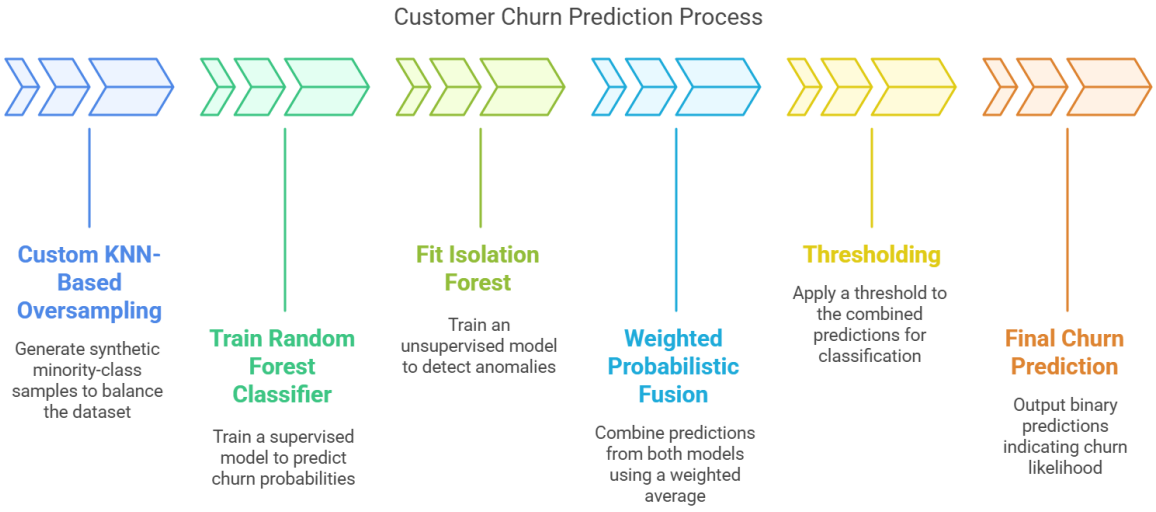

Fig. 1. Churn Prediction Pipeline

trees. This score is normalized using min-max scaling to yield a probabilistic churn proxy:

$$p_{\text{IF}}(x) = \frac{a(x) - \min(a)}{\max(a) - \min(a)} \in [0, 1]$$

This score is particularly valuable in detecting outliers or previously unseen churn behavior that the supervised model may fail to capture.

D. Probabilistic Fusion via Weighted Ensemble

To leverage the strengths of both classifiers, a probabilistic ensemble is constructed using a convex combination of the supervised and unsupervised outputs. The final churn probability for a given sample x is defined as:

$$p_{\text{final}}(x) = \alpha \cdot p_{\text{RF}}(x) + (1 - \alpha) \cdot p_{\text{IF}}(x)$$

where $\alpha \in [0, 1]$ is a tunable fusion coefficient. The optimal value of α is selected via cross-validated grid search to maximize F1-score on the validation fold.

A final binary decision is obtained by thresholding $p_{\text{final}}(x)$ using a data-driven threshold τ^* that maximizes the balance between precision and recall. This ensemble strategy allows the model to adaptively weight reliable historical patterns (captured by Random Forest) and behavioral deviations (captured by Isolation Forest), resulting in enhanced sensitivity to minority-class instances.

IV. Experiments

A. Experimental Setup

To assess the effectiveness of the proposed framework, experiments are conducted on a real-world dataset collected from **Vietcombank**, one of the largest commercial banks in Vietnam. The dataset consists of **10,000 customer records**, each described by 11 structured features including demographic information, account usage patterns, credit behavior, and service activity. The target variable is binary, indicating whether the customer has churned (label = 1) or remains active (label = 0).

Among the 10,000 records, only **2,037** instances correspond to churned customers, resulting in a churn rate of **20.37%**, which poses a significant class imbalance. This makes the dataset well-suited to evaluate methods designed for imbalance-aware classification. Three methods are implemented and compared:

- **Method 1: Autoencoder + Isolation Forest**
  A fully unsupervised baseline that treats churn detection as an anomaly detection task. It combines reconstruction error from an autoencoder with anomaly scores from Isolation Forest.
- **Method 2: ADASYN + Random Forest + IF Ensemble**
  Uses ADASYN to synthetically balance the minority class, trains a Random Forest classifier, and combines its output with Isolation Forest using a rule-based ensemble (logical OR).
- **Method 3 (Proposed): KNNensemble**
  Applies a custom oversampling technique based on K-nearest neighbor interpolation, followed by supervised classification using Random Forest, and final prediction via weighted fusion with Isolation Forest.

B. Results and Discussion

Table I summarizes the classification performance of all three methods. Across all experiments, it is evident that the majority class (no churn) is consistently predicted with high precision, recall, and F1-score across all models. However, the key distinction lies in how well each method handles the minority churn class, where the performance gap becomes substantial.

The **Autoencoder + IF** method performs well on the majority class, with a precision of 0.80 and F1-score of 0.87. However, it fails almost entirely to detect churned customers, achieving only 6% recall and an F1-score of 0.10 for the

TABLE I
PERFORMANCE COMPARISON OF ALL METHODS ON THE TEST SET

Method	Precision (No Churn)	Precision (Churned)	F1 (No Churn)	F1 (Churned)	Recall (No Churn)	Recall (Churned)
Autoencoder + IF	0.80	0.25	0.87	0.10	0.95	0.06
ADASYN + RF + IF	0.88	0.49	0.87	0.52	0.85	0.57
KNNensemble (Ours)	**0.91**	**0.55**	**0.88**	**0.60**	**0.86**	**0.67**

churn class. The **ADASYN + RF + IF** ensemble demonstrates meaningful improvements on the churn class. With a precision of 0.49 and recall of 0.57 for churned customers, it achieves an F1-score of 0.52 — over five times higher than the unsupervised baseline. However, this method is limited by the use of a fixed ensemble rule and ADASYN's tendency to generate borderline or noisy samples.

The proposed KNNensemble model delivers the strongest performance across all metrics. For the churn class, it achieves a precision of 0.55, recall of 0.67, and an F1-score of 0.60, outperforming both baselines. Crucially, it also maintains high precision and F1-score for the majority class (0.91 and 0.88, respectively), demonstrating that it does not sacrifice majority-class accuracy to improve minority-class sensitivity. This balance is achieved through two mechanisms: the generation of locally coherent synthetic churn instances, and the weighted ensemble design, where the fusion weight is tuned to optimize F1-score rather than being hardcoded.

V. CONCLUSION

The proposed KNNensemble framework demonstrates strong potential for addressing class imbalance in churn prediction by effectively combining structure-aware oversampling with adaptive ensemble learning. Without compromising performance on the majority class, the method achieves notable gains in identifying churned customers, making it a practical solution for real-world applications. Future directions include enhancing model interpretability, incorporating temporal behavior, and evaluating real-time deployment in operational systems.

REFERENCES

[1] Adnan Idris, Asifullah Khan, and Yeon Soo Lee. "Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification". In: *Applied Intelligence* 39 (Oct. 2013). DOI: 10.1007/s10489-013-0440-x.

[2] Owolabi et al. "Comparative Analysis of Machine Learning Models for Customer Churn Prediction in the U.S. Banking and Financial Services: Economic Impact and Industry-Specific Insights". In: *Journal of Data Analysis and Information Processing* 36 (2024), pp. 388–418. DOI: 10.4236/jdaip.2024.123021.

[3] J. Burez and D. Van den Poel. "Handling class imbalance in customer churn prediction". In: *Expert Systems with Applications* 36.3, Part 1 (2009), pp. 4626–4636. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2008.05.027. URL: https://www.sciencedirect.com/science/article/pii/S0957417408002121.

[4] Kevin W. Bowyer et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *CoRR* abs/1106.1813 (2011). arXiv: 1106.1813. URL: http://arxiv.org/abs/1106.1813.

[5] Haibo He et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008, pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.

[6] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation Forest". In: *2008 Eighth IEEE International Conference on Data Mining*. 2008, pp. 413–422. DOI: 10.1109/ICDM.2008.17.

[7] Hoang Tran, Ngoc Le, and Van-Ho Nguyen. "Customer Churn Prediction in the Banking Sector Using Machine Learning-Based Classification Models". In: *Interdisciplinary Journal of Information* 18 (Feb. 2023), pp. 087–105. DOI: 10.28945/5086.

[8] Bart Lariviere and Dirk Van den Poel. "Predicting Customer Retention and Profitability by Using Random Forests and Regression Forests Techniques". In: *Expert Systems with Applications* 29 (Aug. 2005), pp. 472–484. DOI: 10.1016/j.eswa.2005.04.043.

[9] Chih-Fong Tsai and Yu-Hsin Lu. "Customer churn prediction by hybrid neural networks". In: *Expert Syst. Appl.* 36 (Dec. 2009), pp. 12547–12553. DOI: 10.1016/j.eswa.2009.05.032.

[10] Sajjad Shumaly, Pedram Neysaryan, and Yanhui Guo. "Handling Class Imbalance in Customer Churn Prediction in Telecom Sector Using Sampling Techniques, Bagging and Boosting Trees". In: Oct. 2020, pp. 082–087. DOI: 10.1109/ICCKE50421.2020.9303698.

[11] Yanjie Dong and Xuehua Wang. "A New Over-Sampling Approach: Random-SMOTE for Learning from Imbalanced Data Sets". In: vol. 7091. Dec. 2011, pp. 343–352. ISBN: 978-3-642-25974-6. DOI: 10.1007/978-3-642-25975-3_30.

[12] Irfan Ullah et al. "Churn Prediction in Banking System using K-Means, LOF, and CBLOF". In: July 2019, pp. 1–6. DOI: 10.1109/ICECCE47252.2019.8940667.

[13] Rui Chen et al. "A Hybrid Framework for Class-Imbalanced Classification". In: Sept. 2021, pp. 301–313. ISBN: 978-3-030-85927-5. DOI: 10.1007/978-3-030-85928-2_24.