The International Journal of Engineering and Science (IJES)
|| Volume || 13 || Issue || 3 || Pages || PP 139-148 || 2024 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Enhancing Customer Churn Prediction: Addressing Disparities and Imbalance in Machine Learning Models.

Emmanuel Chai[1*],Kennedy Khadullo[2] and Kevin Tole[3*]
[1, 2, 3] *Technical University of Mombasa, Institute of Computing and Informatics.*
* *Corresponding author: emarnuelchai@gmail.com; ktole@tum.ac.ke*

----------***ABSTRACT***----------
In this paper we address the Customer Churn Prediction Problem (CCPP). CCPP is an NP-Hard classification problem that involves predicting and identifying potential churners based on historical data so that appropriate measures, interventions, or strategies can be implemented to retain or mitigate the negative impact of their departure. Our objective is to identify and scrutinize methodologies that augment the accuracy and efficacy of customer churn predictions. Through our analysis, it was revealed that automatic churn prediction encounters a challenge stemming from the inherent disparities within the dataset. Specifically, there exists a notable disproportion between the majority and minority classes, potentially leading to model bias that favors the dominant class. This synthesized literature will highlight gaps and limitations in existing research. In conclusion, this literature review presents a significant contribution towards illuminating the need for imbalance correction, thereby fostering an enhanced accuracy and facilitating the progression of future studies.
Keywords*: Customer Churn Prediction, Imbalance Class, Insurance, Machine Learning models.*

Date of Submission: 12-03-2024 Date of acceptance: 27-03-2024

I. Introduction

Given a set of features $\boldsymbol{x}$ represents customer characteristics and behaviors, the goal is to predict the binary outcome y indicating whether a customer churn or not. Let $\boldsymbol{x}$ be a matrix of **size** $\boldsymbol{m\,x\,n}$, where m and n is the total number of samples (customers) and features, respectively. Every column in $\boldsymbol{x}$ represents a feature and every row in $\boldsymbol{x}$ represents a client. Additionally, let y be a binary vector of size $\boldsymbol{m}$, where = 1 $\boldsymbol{y^{I}}$ = 1 if customer i churns and $\boldsymbol{y^{I}}$ = 0 otherwise.

In churn prediction, we aim to learn a function $\boldsymbol{F(x)}$ that maps the feature matrix x to the binary outcome y. This function can be represented by a classifier, such as logistic regression, decision trees, support vector machines, or neural networks. Mathematically, this can be expressed as:

$$\boldsymbol{Y = f(X) + \epsilon} \qquad 1$$

Where ∈ represents the error term.

In this study the NP-Hardness of churn prediction arises from the combinatorial nature of the problem, especially when dealing with large datasets and complex feature spaces. The task involves searching through a vast space of possible feature combinations and model parameters to find the optimal solution that accurately predicts churn. Formally, if we consider N as the number of possible subsets of features, the problem of finding the optimal feature subset for churn prediction becomes combinatorial. This means that the time required to find the optimal solution grows exponentially with the number of features, making it computationally infeasible for large datasets. Moreover, churn prediction often involves dealing with imbalanced datasets, where the number of churners (positive class) is significantly smaller than the number of non-churners (negative class). This further complicates the problem, as standard classification algorithms may struggle to learn from imbalanced data, leading to biased predictions.

Customer churn prediction (CCP) has become a critical task for businesses across various industries, aiming to anticipate and mitigate customer attrition effectively. The advent of machine learning (ML) techniques has significantly enhanced the accuracy and efficiency of churn prediction models. However, challenges persist in the form of disparities and imbalances within the datasets used to train these models. In this introduction, we provide an overview of customer churn prediction, its application areas, variants of machine learning models utilized, previous research works, and outline the contributions of this article.

The application of customer churn prediction spans across various industries, including telecommunications (Eria & Marikannan, 2018), banking (Guliyev & Tatoğlu, 2021), e-commerce(Xiahou & Harada, 2022), subscription services (Katelaris & Themistocleous, 2017), and more. According to (Eria & Marikannan, 2018) in telecommunications, for instance, predicting customer churn helps providers identify

customers at risk of switching to competitors, allowing them to offer targeted retention incentives. Similarly, in e-commerce (Xiahou & Harada, 2022) churn prediction enables businesses to identify customers likely to stop making purchases, allowing them to personalize marketing strategies and improve customer retention.

A wide range of machine learning models are employed in customer churn prediction (CCP), each having its own set of attributes and drawbacks. Commonly used models comprise logistic regression (Jain et al., 2020), decision trees (Bin et al., 2007), random forests, support vector machines(SVM), neural networks, and gradient boosting algorithms (De Caigny et al., 2018). Each model utilizes different algorithms and techniques to analyze data and make predictions, offering varying levels of accuracy and interpretability.

Numerous studies have investigated the effectiveness of different churn prediction models across various industries. Identification and Analysis of Disparities: We identify common disparities and imbalances present in churn prediction datasets, including class imbalance, feature distribution discrepancies, and biases in data collection methods. The following are the major contributions of our work:

I. Identification of Disparities and Imbalances: We identify common disparities and imbalances present in churn prediction datasets, including class imbalance, feature distribution discrepancies, and biases in data collection methods. By recognizing these issues, we lay the foundation for addressing them to improve the performance and fairness of churn prediction models.

II. Evaluation of Previous Work: We review and analyze existing research works in customer churn prediction, highlighting their similarities and differences in addressing disparities and imbalances in ML models.

III. Comparative Analysis: We conduct a comparative analysis of different approaches and techniques proposed in previous research works, assessing their effectiveness in enhancing churn prediction performance.

In this work as shown in Fig.1, content analysis was used to methodically review and analyze already published materials. Extracting significant insights from a wide range of academic publications and research papers was done in an organized and methodical manner. To better understand the discrepancies and imbalance in customer churn prediction models, we aimed to identify the reoccurring themes, patterns, and trends by classifying, tagging, and evaluating the content of these sources. Search engines implemented were Google scholar and Science Direct.

In conclusion, prediction of churn can be framed as a classification problem (Vafeiadis et al., 2015), aiming to learn a function that effectively predicts whether a customer will leave depending on their attributes and habits. However, the NP-Hardness of the problem stems from its combinatorial nature and the challenges posed by imbalanced datasets. Addressing these complexities requires the use of efficient algorithms and optimization techniques to enhance the accuracy of churn prediction models.

To this end, our work is structured as follows: section **2** critically reviews and critiques existing literature. While In section **3**, further discussion on the disparities and imbalances is provided, followed by recommendations in section **4**. Key findings are presented and the research paper concludes in section **5** with conclusive remarks.

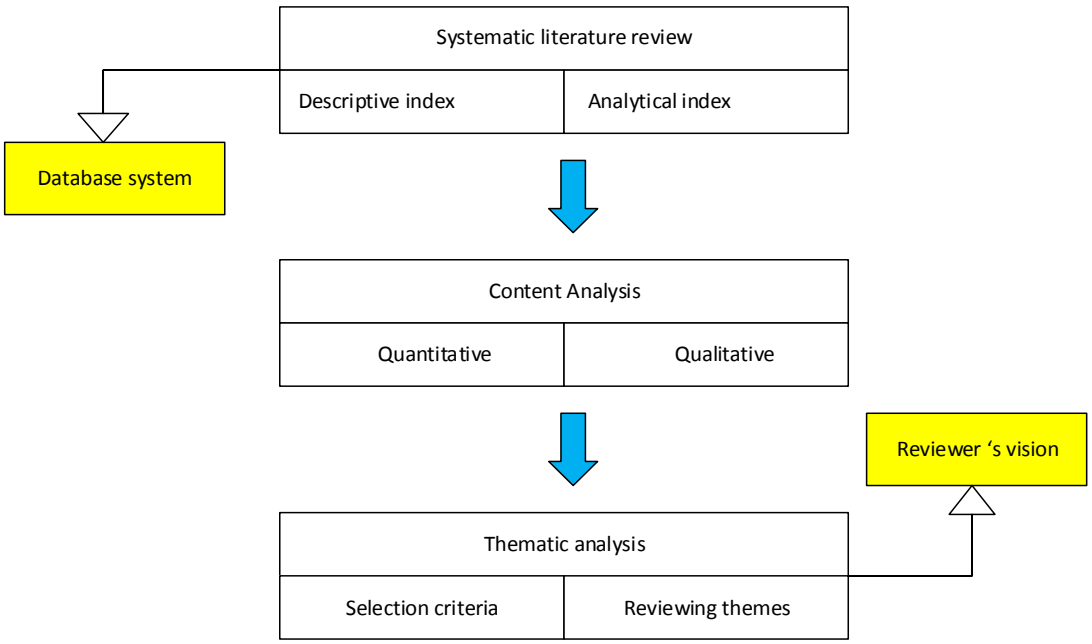

Figure 1. Content Analysis

II. Related Works

This section provides a comprehensive literature review comparing and contrasting findings for more than thirty studies in the field. It emphasizes differences, similarities, and proposes solutions to mitigate methodological challenges. In this study we have structured this section into subsections to facilitate a thorough investigations of the reviews.

2.1 Ensemble Learning

Ensemble learning, which combines multiple models to enhance predictive performance, robustness, and generalization capabilities, involves training diverse base models and aggregating their predictions to make more precise and stable predictions. Ensemble methods, including bagging, stacking, boosting, and hybrid approaches, offer unique advantages in churn prediction tasks.

In their study, (Prashanth et al., 2017) proposed a novel ensemble approach combining random forests, gradient boosting, and Neural Networks for prediction of churn in the telecommunications industry. They demonstrated that ensemble methods outperformed individual models by capturing diverse patterns in the data and reducing overfitting. Conversely, (Bahnsen et al., 2015a) focused on applying ensemble learning with cost-sensitive techniques to address class imbalance in churn prediction for the subscription sector. They utilized a combination of cost-sensitive SVM, decision trees, and neural networks to mitigate biases caused by imbalanced data distribution.

While (Prashanth et al., 2017) emphasized improving overall predictive performance through the combination of diverse models, (Bahnsen et al., 2015a) specifically targeted class imbalance by employing cost-sensitive ensemble methods. Prashanth et al. employed random forests, gradient boosting, and neural networks, while Bahnsen et al . utilized SVM, decision trees, and neural networks with cost-sensitive learning. Both studies underscored the importance of ensemble learning in enhancing churn prediction accuracy and robustness, recognizing the need to leverage multiple models to capture different aspects of churn behavior and mitigate challenges such as overfitting and class imbalance.

In another study, (Abbasimehr et al., 2014) explored a hybrid ensemble approach incorporating bagging and boosting techniques for churn prediction. They demonstrated that combining predictions from bagged and boosted models resulted in improved generalization performance and model stability. Conversely, (Wang et al., 2020) proposed a stacking-based ensemble framework that integrated predictions of base models trained on distinct portions of the data. Their approach leveraged meta-learners to combine diverse predictions and adaptively learn the optimal combination strategy. Smith et al. focused on combining bagging and boosting techniques directly, while Wang et al. employed a stacking-based approach with meta-learners. Smith et al. emphasized the integration of predictions from diverse models, while Wang et al. prioritized adaptive learning of combination strategies.

Both studies highlighted the benefits of ensemble learning in churn prediction, particularly in improving model generalization and stability, recognizing the importance of combining diverse predictions to harness the complementary strengths of individual models and enhance overall predictive performance.

2.2 Machine Learning-Based Approaches

Several studies have focused on evaluating logistic regression, decision trees, and support vector machines (SVM) for predicting customer churn in the telecommunications sector. In comparison, (Abbasimehr et al., 2014) delved into the effectiveness of ensemble methods, specifically Artificial Neural Network and gradient boosting, in predicting churn . While the former observed SVM's superior accuracy but lower interpretability, the latter highlighted boosting had higher accuracy despite biases due to class imbalance. For instance, the study compared the performance of bagging , boosting, staking and voting using four commonly used base learners ANN, SVM, and RIPPER as base learners.. They found that while neural networks achieved the highest accuracy, boosting exhibited better results and computational efficiency. Similarly, In (Coussement & Van den Poel, 2008), used support vector machine (SVM) in a subscription context in order to construct a churn model with a higher predictive performance. The author benchmarked to logistic regression and random forest. Support vector machine showed a good generalization performance when parameters are optimized. When optimal parameters are utilized support vector machine outperforms the logistic regression, whereas random forest outperform both kinds of support vector machines.

Further studies explored various approaches to address specific challenges in churn prediction. (Zhu et al., 2018) compared resampling methods (SMOTE and DBN) with cost-sensitive learning methods (focal loss and weighted loss), finding the latter to have great efficiency. (Zhu et al., 2018) examined the application of decision trees to churn prediction, recommending optimal sampling strategies. In a separate study, (Zhu et al., 2018) benchmarked various sampling techniques, highlighting the importance of evaluation metrics and classifier impact. (idris et al., 2012) suggested an under-sampling technique based on PSO in combination with

feature reduction techniques and Random Forest, achieving strong performance in predicting churners in the telecom industry. These studies collectively underscore the importance of considering both resampling and cost-sensitive learning methods, as well as the need for tailored approaches in specific industry contexts. (R. Zhang et al., 2017) focused on data preprocessing techniques, emphasizing their role in improving model convergence and generalization.

Other studies delved into fairness-aware algorithms. The efficiency of these strategies was shown by (Song et al., 2006) and (Rodan et al., 2015), who both did exceptionally well in field tests and proved that an ensemble of Multilayer perceptrons utilizing negative correlation learning beat other methods. (Dhangar & Anand, n.d. 2021) drew attention to the possibility of biased model training, especially when it comes to comprehensibility and accuracy. (Abbasimehr et al., 2011) underlined this further when she discovered that some methods, including additive logistic regression, may balance comprehensibility and accuracy in churn prediction. All of these researches highlight how important it is to give model training and evaluation significant thought when predicting churn.

Lastly, studies by Xu et al. (2017) and Wu et al. (2017) explored model interpretability techniques and hybrid approaches combining machine learning with expert rules, respectively, to enhance model transparency and performance

2.3 Rule-Based Systems

Rule-based systems provide a transparent and interpretable method for churn prediction, enabling companies to comprehend the factors driving customer attrition. This literature review delves into the conceptual frameworks of ten articles centered on rule-based systems for churn prediction within the telecommunications sector, aiming to delineate disparities and commonalities to elucidate the strengths and weaknesses of varied approaches.

(Y. Huang et al., 2011) introduced the CRL algorithm, which achieved high accuracy and outperformed existing methods. Building on this, (Haghighi et al., 2016) developed a Fuzzy Association Rule-based Classification Learning Algorithm, which also demonstrated high prediction accuracy.

(Amin et al., 2015) In order to choose the most pertinent subset of characteristics, the study employs a three-phase technique that uses a supervised feature selection procedure. By reducing redundancy and boosting relevance, this results in a collection of features that is highly correlated and decreased. A knowledge-based system that uses a ripple down rule learner to gather information about observed customer churn behavior has been developed. Prudence analysis is used to address the issue of brittleness in churn KBS and notifies the decision maker when a case exceeds the knowledge base's-maintained knowledge. In the KB system, Knowledge Acquisition (KA) was assessed using the third simulated expert (SE).Although both (Y. Huang et al., 2011) (Amin et al., 2015) employed rule-based systems for churn prediction, they diverged in their methodologies. While (Y. Huang et al., 2011) relied on predefined rules gleaned from expert knowledge.

On a different note, (B. Huang et al., 2016) presented a classification learning method for churn prediction based on fuzzy association rules. His study on conceptual framework entailed employing rule-based model to generate initial predictions, subsequently refined using a machine learning algorithm. The study aimed to leverage the transparency of rule-based systems while harnessing the predictive power of machine learning models.

Conversely, (Amin et al., 2015) introduced a dynamic rule-based system for churn prediction, continuously updating rules based on evolving customer behavior. Their framework involved monitoring real-time customer interactions and adjusting prediction rules accordingly. The study underscored the adaptability of rule-based systems to dynamic business environments.

While (B. Huang et al., 2016) and (Amin et al., 2015) both proposed innovative strategies to enhance rule-based systems for churn prediction, they diverged in their approaches. (Y. Huang et al., 2011) focused on amalgamating rule-based and machine learning models for augmented accuracy, whereas (Amin et al., 2015) prioritized the dynamic adaptation of rules based on real-time data. These studies highlight how adaptable rule-based systems are to the changing needs of churn prediction in the telecom industry.

2.4 Hybrid Approaches

Hybrid approaches offer a promising avenue for enhancing churn prediction accuracy and addressing specific challenges such as class imbalance and temporal dynamics. While studies vary in their conceptual frameworks and methodologies, they collectively highlight the importance of integrating diverse techniques to develop robust churn prediction models. Further research exploring novel hybridization strategies and evaluating their performance across different industries is warranted to advance the field of churn prediction.

(Bahmani et al., 2013) proposed a hybrid approach integrating cox regression and Neural Networks (ANNs) for prediction of churn in the telecommunications sector. The study combines the interpretability of logistic regression with the nonlinear modeling capabilities of ANNs to improve predictive performance.

A number of studies have explored the potential of hybrid models combining decision trees and support vector machines (SVMs). (Arun Kumar & Gopal, 2010) suggested a hybrid SVM-based decision tree that uses decision trees to roughly represent the SVM decision boundary, , obtaining a notable speedup without sacrificing accuracy. (Kim, 2016)further improved this approach by introducing a semi-supervised decision tree that considers the topological properties of the dataset, leading to superior performance. In the context of regression analysis, (Kim & Hong, 2017) developed a hybrid algorithm that combines a decision tree with a regression algorithm, achieving better or comparable accuracy without significantly increasing computational complexity.

A range of studies have explored the use of hybrid models for churn , with an emphasis on merging survival analysis and random forests. (J. Zhang, 2023) proposed a hybrid RF-MLP algorithm, achieving a high AUC score of 84.9% in customer churn prediction. (Xie et al., 2009a) improved the performance of random forests by integrating a sampling technique and cost-sensitive learning, demonstrating significant accuracy improvements in a credit debt customer database. (Win & Vung, 2023) presented churn prediction models using Gradient Boosted Tree and Random Forest classifiers, achieving high accuracies of 96.2% and 96.89% respectively. (Hastie et al., 2009) introduced random forests as a robust and effective method for classification and regression tasks. (Xie et al., 2009b) further improved this approach by proposing the use of improved balanced random forests (IBRF) for churn prediction, which demonstrated superior performance compared to other algorithms. (Hudaib et al., 2015) explored the use of hybrid models for churn prediction, combining clustering and prediction phases, and found that these models outperformed single common models. (Biau, n.d.2010) provided a statistical analysis of the random forests model, showing its consistency and ability to adapt to sparsity. These studies collectively highlight the potential of hybrid models, particularly those integrating survival analysis and random forests, for improving churn prediction accuracy.

In summary, while both (J. Zhang, 2023)and Gupta et al. propose hybrid approaches for churn prediction, they differ in their methodologies. Kim et al. focus on integrating survival analysis to capture temporal dynamics, whereas Gupta et al. prioritize feature engineering using autoencoders.

III. Disparities and Data Imbalance.

Addressing disparities and imbalances in datasets is essential for enhancing the accuracy and reliability of churn prediction models. To provide our readers with a clearer understanding, in this study we present a mathematical formulation of this concept:

Let $\boldsymbol{D}$ represent the dataset containing churn-related features and target labels. Let $\boldsymbol{x}$ be the feature matrix containing n samples and m features: =[1,2,...,] $\boldsymbol{x}$=**[** $\boldsymbol{x^1}, \boldsymbol{x^2}, \dots, \boldsymbol{x^M}$**]**. Let y be the corresponding target vector indicating churn or non-churn for each sample: =[1,2,...,] $\boldsymbol{y} = [\boldsymbol{y^1}, \boldsymbol{y^2}, \dots . \boldsymbol{y^N}]$, where y^1 is the churn label for sample i. Let wi denote the weight assigned to each sample i based on its class (churn or non-churn). Let pi represent the probability of sample I belonging to the churn class. Let $\hat{}p\hat{}i$ be the predicted probability of sample i belonging to the churn class obtained from the churn prediction model. Let λ be the regularization parameter controlling the impact of misclassification costs in the model.

The objective function to address disparities and imbalance in churn prediction can be formulated as:

$$\frac{\boldsymbol{min}}{\boldsymbol{w,p,\lambda}} [\sum_{\boldsymbol{i=0}}^{\boldsymbol{n}} \boldsymbol{w_i}. \boldsymbol{cost}(\boldsymbol{y_i}, \boldsymbol{p_i}) + \boldsymbol{\lambda}. \boldsymbol{\Omega}(\boldsymbol{w})] \quad 2$$

Where:

Cost (yi,$p\hat{}i$) - is the misclassification cost function, penalizing errors differently based on the class distribution.

Ω()Ω(**w**) - is the regularization term to control the complexity of the model and prevent overfitting.

w - is the vector of sample weights assigned to address class imbalance.

P - Represents the vector of probabilities indicating the likelihood of each sample belonging to the churn class.

Several studies have proposed various techniques to mitigate these challenges. (Bahnsen et al., 2015b) proposed a method for addressing class imbalance in churn prediction by incorporating cost-sensitive learning techniques. They adjusted the misclassification costs to penalize errors differently based on the class distribution, thereby reducing the impact of class imbalance.

While (Bahnsen et al., 2015a) approach effectively addressed class imbalance, it primarily focused on adjusting misclassification costs. However, it may not fully capture the underlying complexities of imbalanced data, such as overlapping distributions or rare events. In (Douzas et al., 2018), researchers introduced a synthetic minority oversampling technique (SMOTE) to augment the minority class in churn prediction datasets. They showed that oversampling techniques like SMOTE could enhance model accuracy by generating synthetic instances of minority class samples.

(Douzas et al., 2018) approach is effective in balancing class distributions, but it may lead to overfitting, especially when generating synthetic samples that do not accurately represent the underlying data distribution. Additionally, SMOTE may not fully capture the intricacies of the original data, potentially affecting model generalization. (Abbasimehr et al., 2014) proposed a hybrid approach combining ensemble learning with data-level resampling techniques for churn prediction. They integrated techniques such as bagging and boosting with under sampling and oversampling to handle class imbalance effectively. (Kimura, 2022) hybrid approach is comprehensive in addressing class imbalance by leveraging both ensemble learning and resampling techniques. However, the computational complexity of ensemble methods may limit scalability, particularly with large datasets.

(Naing et al., 2022) investigated the impact of feature selection on addressing disparities in churn prediction. They identified informative features using various selection algorithms, such as mutual information and recursive feature elimination, to improve model performance. They focus on feature selection is valuable in reducing dimensionality and improving model interpretability. However, feature selection alone may not fully mitigate class imbalance or address disparities in the underlying data distribution.

(Chawla et al., 2002) proposed a framework for ensemble-based anomaly detection to identify rare churn events in imbalanced datasets. They combined multiple anomaly detection algorithms to detect unusual patterns indicative of churn. Their approach to anomaly detection offers a unique perspective on addressing class imbalance by identifying rare churn events. However, it may overlook the broader context of churn prediction and may not be suitable for datasets with high-dimensional feature spaces.

In summary, while existing literature offers diverse approaches to address disparities and imbalance in churn prediction datasets, each method has its strengths and limitations. Future research should focus on integrating multiple techniques and developing hybrid approaches to enhance the robustness and effectiveness of churn prediction models. As shown in **Table 1**, we summarize all churn prediction related work. The first row displays the first Author and year of publication, and row two shows the methodology used in that research. According to the best of our knowledge the remain rows display the advantages, limitations and recommendations from row 3- 5 respectively.

Table 1.Summary of churn prediction related researches

Author & year	Methodologies	Advantages	Limitations	Recommendations
Abbasimehr, hossein (2014)	Bagging, boosting, stacking	Use of ensemble methods	The study does not consider the impact of business objectives on model performance.	Ensemble learning, particularly heterogeneous ensembles, can significantly improve the accuracy and performance.
Abbasimehr, hossein(2011)	Logistic regression (lr) and additive logistic regression (alr)	Use of combinational method of divide and conquer and separate and conquer algorithms	Lack of real-world application and the need for further validation of the proposed methods in practical settings.	Use of confusion matrix and precision for performance measure
Amin, adnan(2015)	Ripple down rule (rdr), simulated expert (se)	Prudent and simulation expert-based approach in combination with rdr classifier resulted in promising results.	Several researchers have used various algorithms and strategies to approach the problem of client churn. .	Combination of feature selection, deep learning, customer segmentation, and machine learning models can enhance the accuracy of customer churn prediction.
Arun kumar(2010)	Svm	Hybridize svm and dt to exploit the advantages of both the learning methodologies	The study focused on reducing the number of test datapoints that need svm's decision in getting classified.	Use of a hybrid svm-based decision tree for improved classification performance and reduced computational complexity.
Bahnsen, alejandro correa(2013)	Neural network, cox regression	Proposed model outperformed pure methods	Overfitting in the random forest model, as well as the poor convergence and lower efficiency of feature selection techniques	Combination of random forest models, customer segmentation, and explainable machine learning methods
Bahnsen, a. C., aouada, d., & ottersten, b. (2015a)	Cost-sensitive framework	Emphasize on the importance of a cost-sensitive approach by demonstrating a 26.4% increase in cost savings	Lack of detailed discussion of the limitations of the proposed framework, particularly in terms of its applicability to different industries and the potential biases in the cost optimization process	Need for a cost-sensitive approach in customer churn prediction and the potential of machine learning and data mining techniques in this domain.
Bin, l., peiji, s., &	Decision tree	Use of optimal parameters for	Lack of customer	Use of decision trees with

juan, l. (2007)		the model, such as sub-period length, misclassification cost, and sample method, to enhance its performance	information and skewed class distribution	specific parameters such as sub-period length, misclassification cost, and sample method.
Chawla, n. V., bowyer, k. W., hall, l. O., & kegelmeyer, w. P. (2002)	Under-sampling , over-sampling minority class	Enhancing of smote to address its limitations	Overfitting in smote	Use of generative adversarial network to improve classifier performance.
Coussement, k., & van den poel, d. (2008)	Support vector machine	SVM outperformed traditional logistic regression, with the parameter-selection technique playing a crucial role in predictive performance	Comparison of parameter-selection techniques may not be exhaustive, and the study's findings may not be generalizable to other industries or contexts	Use of artificial neural network and gaussian naïve bayes
De caigny, a., coussement, k., & de bock, k. W. (2018).	Logistic regression and decision trees	Hybrid classification algorithm outperforms both decision trees and logistic regression in terms of predictive performance and comprehensibility	While the llm outperforms its building blocks in terms of predictive performance, it still faces limitations in handling linear relations between variables and interaction effects	Have a focus on reducing execution time.
Dieng, a. B., ruiz, f. J. R., blei, d. M., & titsias, m. K. (2019)	Gans	Useful in optimization function, metrics, and implementation,	They may be more effective in smaller domain shifts/dataset	Emphasizing the role in generating realistic adversarial data.
Douzas, g., bacao, f., & last, f. (2018)	K-means and smote	Use of k-means clustering, helps avoid the generation of unnecessary noise.	Not considerate on the data distribution and density information, which are crucial for synthesizing minority examples	Focus on oversampling methods based on k-means and smote, particularly those that address data density and class imbalance
Eria, k., & marikannan, b. P. (2018)	Decision tree, support vector, naïve bayes	Highlight the importance of machine learning techniques, proposing a hybrid approach	Lack of in-depth analysis of the specific models and methods used in churn prediction, and the absence of a comprehensive comparison of these models.	Prevalence of data mining methods, such as artificial neural network, in this area.
Guliyev, h., & tatoğlu, f. Y. (2021)	Xgboost	Xgboost outperformed other machine learning approaches in classifying churn clients.	Does not address the potential biases in the data that could affect the model's performance	Combination of random forest models, customer segmentation, and explainable machine learning methods
Huang, b., huang, y., chen, c., & kechadi, m.-t. (2016)	Caim discretization algorithm	Proposed rule-based classification algorithms, with the former achieving acceptable prediction accuracy and efficiency, and the latter outperforming existing methods	disregards the variables that are rich in information found in call detail records (CDRs)	Emphasize the possibilities of rule-based and fuzzy logic approaches for predicting client attrition. .
Jain, h., khunteta, a., & srivastava, s. (2020).	Logistic regression and logit boost	Provides a robust predictive model for identifying potential churners, enhancing precision and efficiency	Could not consider the potential benefits of boosting algorithms in separating high-risk customer clusters	Emphasizing the model's interpretability and predictive power.

IV. Key Findings and Recommendations.

Addressing disparities and imbalances in datasets is pivotal for enhancing the predictability and reliability of predicting churn models. The key findings of addressing disparities and imbalance in churn prediction models revolve around the effectiveness of various techniques in mitigating these challenges. These findings are crucial for improving the accuracy and reliability of churn prediction systems.

One significant finding is the effectiveness of oversampling techniques, such as Synthetic Minority Oversampling Technique (SMOTE), in balancing class distributions in imbalanced datasets. By generating synthetic instances of minority class samples, SMOTE helps in improving model performance and reducing bias towards majority classes. Hybrid approaches, which combine multiple techniques such as ensemble learning, feature selection, and cost-sensitive learning, have been found to be effective in addressing disparities and imbalances.

These approaches build on the qualities of many methodologies to improve forecast accuracy and robustness. Feature engineering is critical in addressing discrepancies by discovering informative features and

lowering dimensionality. Techniques such as recursive feature elimination and mutual information have been shown to improve model performance by selecting relevant features and reducing noise in the data.

While oversampling techniques can effectively balance class distributions, they may lead to overfitting, especially when generating synthetic samples that do not accurately represent the underlying data distribution. Moreover, extreme class imbalances can result in biased predictions, highlighting the need for balanced sampling strategies.

The importance of model interpretability is also emphasized in churn prediction. While improving prediction accuracy is essential, model interpretability is crucial for understanding the factors driving churn. Interpretable models provide insights into the decision-making process, especially in domains where regulatory compliance and transparency are critical. Some approaches, such as ensemble methods and hybrid models, may face challenges in scalability and computational complexity, particularly with large datasets. Balancing model performance with computational efficiency is essential for practical deployment in real-world scenarios. While many studies focus on specific industries such as telecommunications and banking, generalizing findings across different sectors remains a challenge. Factors such as data characteristics, business dynamics, and customer behaviors vary across industries, necessitating tailored approaches for churn prediction. Overall, the key findings underscore the importance of adopting a multifaceted approach to address disparities and imbalances in churn prediction models. By leveraging a combination of oversampling techniques, hybrid approaches, feature engineering, and interpretable models, researchers and practitioners can develop more accurate and reliable churn prediction systems. However, ongoing research is needed to explore new methodologies and evaluate their effectiveness in diverse real-world settings. To bridge this gap, several strategies can be employed for instance:

a) Class Imbalance Techniques: Employ techniques such as oversampling, under-sampling, or hybrid approaches to balance the class distribution in the dataset. Oversampling techniques generate synthetic samples for the minority class, while under-sampling reduces the majority class instances. Hybrid approaches combine oversampling and under-sampling to achieve a more balanced dataset.

b) Cost-sensitive Learning: Use cost-sensitive learning methods to apply varying misclassification costs to distinct classes based on distribution. By penalizing errors differently, these algorithms help mitigate the impact of class imbalance and disparities.

c) Feature Engineering: Carefully select and engineer features that are informative and relevant for churn prediction. Feature selection methods can help to minimize complexity and focus on the most discriminative attributes, thereby improving model performance and generalization.

d) Model Evaluation Metrics: Use appropriate evaluation metrics that account for class imbalance, such as precision, recall, F1-score, or area under the receiver operating characteristic curve (AUC-ROC). These metrics provide a more comprehensive assessment of model performance, especially in imbalanced datasets.

e) Ensemble Methods: To reduce biases and enhance overall performance, use ensemble methods to aggregate predictions from different models. Ensemble techniques such as bagging, boosting, or stacking can effectively address disparities and imbalances by leveraging the strengths of different base models.

f) Regularization Techniques: Apply regularization techniques to prevent overfitting and improve model generalization. Regularization penalizes overly complex models, thereby reducing the risk of capturing noise or biases present in imbalanced datasets.

g) Fairness-aware Algorithms: Develop fairness-aware algorithms that explicitly address disparities and biases in predictions across different demographic groups. These algorithms aim to minimize disparate impacts and ensure fairness in model predictions.

By integrating these strategies, the gap in addressing disparities and imbalances in datasets can be effectively bridged, leading to more precise and dependable churn prediction models. However, these approaches must be rigorously evaluated and validated to ensure their efficacy and resilience across a wide range of datasets and applications. Additionally, ongoing research and collaboration between academia and industry are crucial for advancing the field and developing innovative solutions to tackle this challenge.

V. Conclusion

Customer churn prediction is an essential responsibility for organizations in variety industries, aiming to anticipate and mitigate customer attrition effectively. This task requires learning a function that accurately predicts if a client will churn based on their features and actions. However, the NP-Hardness of the problem arises from its combinatorial nature, especially when dealing with large datasets and complex feature spaces. Moreover, challenges such as imbalanced datasets further complicate the problem, as standard classification algorithms may struggle to learn from imbalanced data, leading to biased predictions. While churn prediction has been achieved through the use of several machine learning algorithms, each with its unique strengths and limitations, Future research in churn prediction could focus on algorithmic improvements tailored for handling imbalanced datasets and the combinatorial nature of churn prediction, aiming to develop more efficient

algorithms and optimization techniques. Additionally, investigating novel feature engineering techniques could enhance the extraction of informative features from the data, thereby improving model performance. Standardized evaluation metrics and protocols are needed to assess churn prediction models accurately, particularly in the presence of imbalanced datasets, highlighting the importance of model evaluation. Furthermore, addressing biases in data collection methods and mitigating biases in models is crucial for ensuring fairness in churn prediction, especially in sensitive domains such as banking and telecommunications, emphasizing the need for fairness and bias mitigation strategies in future research endeavors.

REFERENCES

[1]. Abbasimehr, H., Setak, M., & Tarokh, M. J. (2014). A comparative assessment of the performance of ensemble learning in customer churn prediction. Int. Arab J. Inf. Technol., 11(6), 599–606.

[2]. Abbasimehr, H., Tarokh, M. J., & Setak, M. (2011). Determination of Algorithms Making Balance Between Accuracy and Comprehensibility in Churn Prediction Setting. International Journal of Information Retrieval Research (IJIRR), 1(2), 39–54. https://doi.org/10.4018/ijirr.2011040103

[3]. Amin, A., Rahim, F., Ramzan, M., & Anwar, S. (2015). A Prudent Based Approach for Customer Churn Prediction. In S. Kozielski, D. Mrozek, P. Kasprowski, B. Małysiak-Mrozek, & D. Kostrzewa (Eds.), Beyond Databases, Architectures and Structures (Vol. 521, pp. 320–332). Springer International Publishing. https://doi.org/10.1007/978-3-319-18422-7_29

[4]. Arun Kumar, M., & Gopal, M. (2010). A hybrid SVM based decision tree. Pattern Recognition, 43(12), 3977–3987. https://doi.org/10.1016/j.patcog.2010.06.010

[5]. Bahmani, B., Mohammadi, G., Mohammadi, M., & Tavakkoli-Moghaddam, R. (2013). Customer churn prediction using a hybrid method and censored data. Management Science Letters, 3(5), 1345–1352.

[6]. Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015a). A novel cost-sensitive framework for customer churn predictive modeling. Decision Analytics, 2(1), 5. https://doi.org/10.1186/s40165-015-0014-6

[7]. Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015b). A novel cost-sensitive framework for customer churn predictive modeling. Decision Analytics, 2(1), 5. https://doi.org/10.1186/s40165-015-0014-6

[8]. Biau, G. (n.d.). Analysis of a Random Forests Model.

[9]. Bin, L., Peiji, S., & Juan, L. (2007). Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service. 2007 International Conference on Service Systems and Service Management, 1–5. https://doi.org/10.1109/ICSSSM.2007.4280145

[10]. 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.

[11]. Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), 313–327.

[12]. De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2), 760–772. https://doi.org/10.1016/j.ejor.2018.02.009

[13]. Dhangar, K., & Anand, P. (n.d.). A REVIEW ON CUSTOMER CHURN PREDICTION USING MACHINE LEARNING APPROACH. 8(5).

[14]. Dieng, A. B., Ruiz, F. J. R., Blei, D. M., & Titsias, M. K. (2019). Prescribed Generative Adversarial Networks (arXiv:1910.04302). arXiv. http://arxiv.org/abs/1910.04302

[15]. Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information Sciences, 465, 1–20. https://doi.org/10.1016/j.ins.2018.06.056

[16]. Eria, K., & Marikannan, B. P. (2018). Systematic review of customer churn prediction in the telecom sector. Journal of Applied Technology and Innovation, 2(1).

[17]. Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. Journal Of Applied Microeconometrics, 1(2), 85–99.

[18]. Haghighi, M., Johnson, S. B., Qian, X., Lynch, K. F., Vehik, K., & Huang, S. (2016). A Comparison of Rule-based Analysis with Regression Methods in Understanding the Risk Factors for Study Withdrawal in a Pediatric Study. Scientific Reports, 6(1), 30828. https://doi.org/10.1038/srep30828

[19]. Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random Forests. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction (pp. 587–604). Springer. https://doi.org/10.1007/978-0-387-84858-7_15

[20]. Huang, B., Huang, Y., Chen, C., & Kechadi, M.-T. (2016). A Fuzzy Rule-Based Learning Algorithm for Customer Churn Prediction. In P. Perner (Ed.), Advances in Data Mining. Applications and Theoretical Aspects (pp. 183–196). Springer International Publishing. https://doi.org/10.1007/978-3-319-41561-1_14

[21]. Huang, Y., Huang, B., & Kechadi, M.-T. (2011). A Rule-Based Method for Customer Churn Prediction in Telecommunication Services. In J. Z. Huang, L. Cao, & J. Srivastava (Eds.), Advances in Knowledge Discovery and Data Mining (pp. 411–422). Springer. https://doi.org/10.1007/978-3-642-20841-6_34

[22]. Hudaib, A., Dannoun, R., Harfoushi, O., Obiedat, R., & Faris, H. (2015). Hybrid data mining models for predicting customer churn. International Journal of Communications, Network and System Sciences, 8(5), 91–96.

[23]. Idris, A., Rizwan, M., & Khan, A. (2012). Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies. Computers & Electrical Engineering, 38(6), 1808–1819. https://doi.org/10.1016/j.compeleceng.2012.09.001

[24]. Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn Prediction in Telecommunication using Logistic Regression and Logit Boost. Procedia Computer Science, 167, 101–112. https://doi.org/10.1016/j.procs.2020.03.187

[25]. Kate, P., Ravi, V., & Gangwar, A. (2022). FinGAN: Generative Adversarial Network for Analytical Customer Relationship Management in Banking and Insurance (arXiv:2201.11486). arXiv. https://doi.org/10.48550/arXiv.2201.11486

[26]. Katelaris, L., & Themistocleous, M. (2017). Predicting Customer Churn: Customer Behavior Forecasting for Subscription-Based Organizations. In M. Themistocleous & V. Morabito (Eds.), Information Systems (pp. 128–135). Springer International Publishing. https://doi.org/10.1007/978-3-319-65930-5_11

[27]. Kim, K. (2016). A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree. Pattern Recognition, 60, 157–163. https://doi.org/10.1016/j.patcog.2016.04.016

[28]. Kim, K., & Hong, J. (2017). A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis. Pattern Recognition Letters, 98, 39–45. https://doi.org/10.1016/j.patrec.2017.08.011
[29]. Kimura, T. (2022). Customer Churn Prediction with Hybrid Resampling and Ensemble Learning. 1–23.
[30]. Naing, Y. T., Raheem, M., & Batcha, N. K. (2022). Feature Selection for Customer Churn Prediction: A Review on the Methods & Techniques applied in the Telecom Industry. 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), 1–5. https://doi.org/10.1109/ICDCECE53908.2022.9793315
[31]. Prashanth, R., Deepak, K., & Meher, A. K. (2017). High Accuracy Predictive Modelling for Customer Churn Prediction in Telecom Industry. In P. Perner (Ed.), Machine Learning and Data Mining in Pattern Recognition (Vol. 10358, pp. 391–402). Springer International Publishing. https://doi.org/10.1007/978-3-319-62416-7_28
[32]. Rodan, A., Fayyoumi, A., Faris, H., Alsakran, J., & Al-Kadi, O. (2015). Negative Correlation Learning for Customer Churn Prediction: A Comparison Study. The Scientific World Journal, 2015, e473283. https://doi.org/10.1155/2015/473283
[33]. Song, G., Yang, D., Wu, L., Wang, T., & Tang, S. (2006). A Mixed Process Neural Network and its Application to Churn Prediction in Mobile Communications. Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06), 798–802. https://doi.org/10.1109/ICDMW.2006.12
[34]. Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. Ch. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1–9. https://doi.org/10.1016/j.simpat.2015.03.003
[35]. Wang, X., Nguyen, K., & Nguyen, B. P. (2020). Churn Prediction using Ensemble Learning. Proceedings of the 4th International Conference on Machine Learning and Soft Computing, 56–60. https://doi.org/10.1145/3380688.3380710
[36]. Win, Y. Y., & Vung, C. G. (2023). Churn Prediction Models Using Gradient Boosted Tree and Random Forest Classifiers. 2023 IEEE Conference on Computer Applications (ICCA), 271–275. https://ieeexplore.ieee.org/abstract/document/10181933/
[37]. Xiahou, X., & Harada, Y. (2022). B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), Article 2. https://doi.org/10.3390/jtaer17020024
[38]. Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009a). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), 5445–5449.
[39]. Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009b). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), 5445–5449.
[40]. Zhang, J. (2023). Customer churn prediction based on a novelty hybrid random forest algorithm. Third International Conference on Computer Vision and Data Mining (ICCVDM 2022), 12511, 623–628. https://doi.org/10.1117/12.2660705
[41]. Zhang, R., Li, W., Tan, W., & Mo, T. (2017). Deep and shallow model for insurance churn prediction service. 2017 IEEE International Conference on Services Computing (SCC), 346–353.
[42]. Zhu, B., Baesens, B., Backiel, A., & vanden Broucke, S. K. L. M. (2018). Benchmarking sampling techniques for imbalance learning in churn prediction. Journal of the Operational Research Society, 69(1), 49–65. https://doi.org/10.1057/s41274-016-0176-1