International Journal of Analysis and Applications

Comparative Analysis of Data Augmentation Methods for Enhancing the Performance of Churn Prediction Models

Faroug A. Abdalla[1], Zakariya M. S. Mohammed[2,3,*], Ali Satty[3], Ashraf F. A. Mahmoud[1,4], Mohamed Ben Ammar[2], Abdelnasser Saber Mohamed[5], Entisar H. Khalifa Osman[1], Shimaa A. Ahmed[6]

[1]*Department of Computer Science, College of Science, Northern Border University, Arar, Saudi Arabia*

[2]*Center for Scientific Research and Entrepreneurship, Northern Border University, 73213, Arar, Saudi Arabia*

[3]*Department of Mathematics, College of Science, Northern Border University, Arar, Saudi Arabia*

[4]*Translation, Authorship, and Publishing of Center, Northern Border University, Arar, Saudi Arabia*

[5]*Computer Science Department, Applied College, Northern Border University, Arar, Saudi Arabia*

[6]*Department of Electrical Engineering, College of Engineering, Northern Border University, Arar, Saudi Arabia*

**Corresponding author:* zakariya.mohammed@nbu.edu.sa

ABSTRACT. Customer churn is a critical challenge for subscription-based businesses, often exacerbated by imbalanced datasets that hinder predictive accuracy. This study evaluates various oversampling techniques, K-means SMOTE, SMOTE, and ADASYN, that generate synthetic samples to balance datasets. The objective is to assess the impact of these oversampling techniques on the performance of machine learning (ML) classifiers, including gradient boosting (GB), random forest (RF), naive Bayes (NB), and support vector machines (SVM). Findings reveal that K-means SMOTE is the most effective at improving model performance, while GB consistently outperforms other classifiers in churn prediction. These findings provide valuable insights into optimizing data balancing and predictive models, offering a robust framework to enhance customer retention strategies.

1. Introduction

Churn prediction plays a vital role in customer relationship management by identifying customers who may soon stop using a company's services or cancel their subscriptions. This is

Received Jul. 31, 2025

2020 Mathematics Subject Classification. 68T05.

Key words and phrases. imbalanced data; churn prediction; data augmentation; SMOTE, k-mean SOMTE; ADASYN.

https://doi.org/10.28924/2291-8639-23-2025-296

ISSN: 2291-8639

especially critical in subscription-based businesses, where retaining existing customers is more cost-effective than acquiring new ones [1]. However, a prevalent issue in churn prediction is the imbalance in datasets, where the number of churners is significantly lower than that of non-churners. This imbalance can lead to classifiers performing well on accuracy but poorly identifying actual churners [2]. Such issues can result in misclassifying loyal customers as churners, which wastes resources, or missing churners altogether, risking customer loss. To tackle this, it is essential to pre-process and balance the data using machine learning (ML) techniques. Businesses can develop more reliable predictive models to support targeted retention strategies by doing so.

Several ML oversampling techniques, such as the minority over-sampling technique (SMOTE), K-means SMOTE, and the adaptive synthetic sampling approach (ADASYN), have been proposed to address the imbalance problem in churn prediction. These techniques are essential for enhancing the predictive accuracy of models identifying potential churners in heavily imbalanced datasets. SMOTE is a widely used technique. Use single line spacing, 3 pt. Spacing after the paragraph. All levels of headings should use 12 pt. spacing before, and 3 pt. Spacing after. After.

Technique that interpolates between existing minority class samples to generate new instances, providing a straightforward yet effective solution [3]. K-means SMOTE combines the clustering capabilities of K-means with SMOTE to create more precise synthetic samples, thereby balancing datasets and improving model performance [4]. Meanwhile, ADASYN further refines this process by concentrating on generating data for more challenging instances, thus enhancing the model's ability to detect churners [5]. This study focuses on the application of these techniques and their effect on the performance of well-known ML classifiers, Gradient Boosting (GB), Random Forest (RF), Naive Bayes (NB), and Support Vector Machine (SVM) in the context of churn prediction.

GB is particularly effective due to its ability to enhance model accuracy by combining multiple weak models into a strong predictive model [6]. RF, known for its robustness in handling diverse datasets, is frequently applied in churn prediction to provide high accuracy and interpretability [7]. NB remains a straightforward yet powerful method, especially in high data dimensions, offering reliable predictions with minimal computational cost [8]. Furthermore, SVM is leveraged for its efficacy in handling high-dimensional data, making it suitable for complex churn prediction tasks [5]. This study provides a comprehensive analysis of optimizing ML classifiers for churn prediction, thereby supporting more effective customer retention strategies by evaluating these classifiers under different oversampling scenarios.

Considering the critical problem of imbalanced data in churn prediction, this research explores how specific oversampling techniques, K-means SMOTE, SMOTE, and ADASYN,

impact the performance of ML classifiers, including GB, RF, NB, and SVM. The objective is to examine these techniques' effects on class imbalance and evaluate the performance of each classifier using various evaluation metrics. The novelty of this study lies in its detailed analysis of the interactions between oversampling techniques and ML classifiers, highlighting the unique responses of each classifier to data balancing. This comprehensive exploration provides valuable insights into optimizing churn prediction models. The importance of this research is underscored by its potential to enhance customer retention strategies through more accurate and tailored ML approaches, benefiting subscription-based businesses by improving predictive accuracy and resource allocation.

2. Methodology

The study process, illustrated in Figure 1, begins with extracting the telecommunication industry customer churn dataset. The data is preprocessed through cleaning and churn prediction using the original data with imbalanced cases. Next, customer churn prediction is conducted on the imbalanced dataset using various ML classifiers. The data is then balanced utilizing oversampling techniques. ML classifiers are then applied to these balanced datasets. Performance is analyzed using multiple evaluation metrics for both imbalanced and balanced datasets.

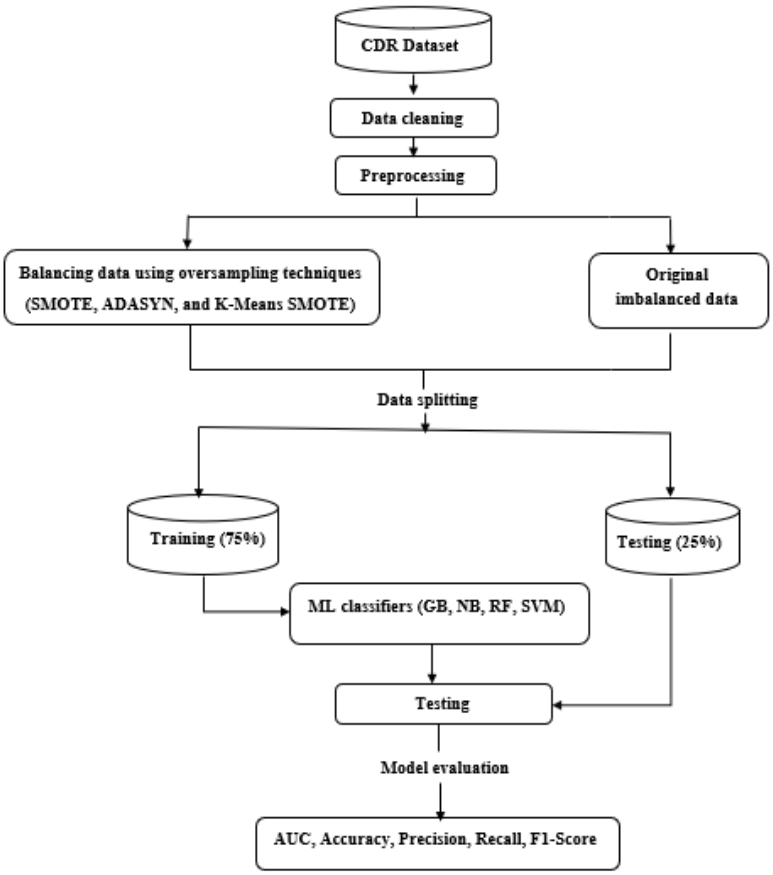


Figure 1. Framework of data preprocessing, balancing, and prediction techniques

2.1 Data source

The data used in this study offers insights into customer behaviors at a telecommunications company. The study sample includes 7,043 records with 21 features, which can be utilized to predict the likelihood of customer churn. The dataset shows that 1,869 customers have churned, while 5,174 customers have not. This data is part of the Call Detail Records file and is available for open access on Kaggle. The study characteristics are outlined in Table 1. More detailed information about this data can be found at [9].

Table 1. Data description of the study characteristics

Features	Data format	Description
Target variable		
Churn	Yes/No	Customer status regarding churn
Predictors		
CustomerID	String	Customer ID
GENDER	Female /Male	Gender of the customer
SeniorCitizen	0/1	Does the customer belong to a senior citizen group
Partner	Yes/No	Does the customer have a partner
Dependents	Yes/No	Do the customers' dependents
Tenure	Numeric	Length of service
PhoneService	Yes/No	Subscription to a phone service
MultipleLines	Yes/No	Subscription to multiple phone lines
InternetService	DSL/Fiber optic /No	Internet service provider
OnlineSecurity	Yes/No	Subscription to online security
OnlineBackup	Yes/No	Subscription to an online backup
DeviceProtection	Yes/No	Subscription to device protection
TechSupport	Yes/No	Subscription to tech support
StreamingTV	Yes/No	Subscription to streaming TV
StreamingMovies	Yes/No	Subscription to streaming movies
Contract	Month-to-month /One-year /Two-year	Type and length of contract
PaperlessBilling	Yes/No	Paperless billing usage
PaymentMethod	Bank transfer /Credit card /Electronic check /Mailed check	Payment method
MonthlyCharges	Numeric	Monthly charge
TotalCharges	Numeric	Total charge

2.2 Data preprocessing

It is a critical step in ML to ensure the dataset is clean, structured, and ready for effective modelling. In this study, missing data were addressed using multiple imputation, a robust method that estimates and fills in missing values to maintain the dataset's integrity. All string and categorical variables were converted into nominal data types, a vital transformation that enhances the accuracy of ML classifiers by ensuring these variables are appropriately processed during training. A significant challenge with the dataset was its class imbalance: the "customers are not churned" class accounted for 73.5%, while the "customers are churned" class comprised 26.5%. This imbalance could bias classifiers towards the majority class, making them more likely to classify new observations as "customers are not churned". To address this issue, the oversampling techniques discussed below were applied to balance the dataset and improve model performance.

2.2.1 Synthetic minority over-sampling technique (SMOTE)

SMOTE, developed by Chawla et al. [10], is one of the pioneering techniques for addressing class imbalance by creating synthetic samples of the minority class. It works by selecting the nearest neighbors of minority class instances and generating new synthetic instances along the line segments joining these neighbors. SMOTE effectively increases the minority class sample size without merely duplicating existing instances, which can lead to overfitting. In the context of churn prediction, it is applied to data where churned customers represent a small fraction of the total data. By enhancing the minority class, SMOTE helps train models that are less biased towards the majority class and, therefore, more capable of identifying potential churners. The formula for SMOTE is as follows:

$$\boldsymbol{x}_{new} = \boldsymbol{x}_i + \delta.(\boldsymbol{x}_{zi} - \boldsymbol{x}_i)$$

Where $\boldsymbol{x}_i$ is a minority class sample, $\boldsymbol{x}_{zi}$is a randomly selected neighbor from the k-nearest neighbors, and $\boldsymbol{\delta} \sim \cup (\mathbf{0}, \mathbf{1})$ is a random number between $\mathbf{0}$ and $\mathbf{1}$.

2.2.2 K-means SMOTE

It is an advanced variation of the SMOTE algorithm that combines clustering with synthetic sampling to address dataset imbalance more effectively [11]. Integrating K-means clustering first divides the dataset into clusters based on the minority class instances. Within each cluster, SMOTE is applied to generate synthetic cases. This localized approach ensures that the synthetic data is more representative of the actual data distribution, considering the natural grouping within the data. K-means SMOTE creates a more balanced and informative training dataset in churn prediction. This is particularly beneficial in scenarios where distinct customer segments influence churn behavior. The predictive models can better understand and anticipate customer churn patterns by ensuring that synthetic samples are generated to reflect these segments. The formula for K-means SMOTE is as follows:

$$\underset{C}{argmin} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

Where k is the number of clusters, C_i is the set of points in cluster i, and μ_i is the centroid of cluster i.

2.2.3 Adaptive synthetic sampling approach (ADASYN)

It is an advanced oversampling technique to handle the issue of class imbalance, which is prevalent in many real-world datasets, including churn prediction scenarios. Introduced by He et al. [12], it focuses on adapting the learning approach according to the data distribution. The core idea is to generate synthetic data points for the minority class by considering the density distribution of the minority class instances. Essentially, more synthetic data is generated for minority class instances that are harder to classify, which is determined by the local distribution of the data. When applied to churn prediction, ADASYN balances the dataset by generating synthetic samples for the minority (churned) class, enabling models trained on such datasets to achieve greater accuracy in predicting churn. The formula for K-means SMOTE is as follows:

$$g_i = \frac{r_i}{\sum_{j=1}^{n_{min}} r_j} \cdot G$$

Where is $r_i = \frac{\Delta i}{k,}$ the ratio of majority class samples in the k-nearest neighbors of x_i, Δ_i is the number of majority class neighbors, G is the total number of synthetic samples to generate, and n_{min} is the number of minority samples

2.3 Model building

The dataset was randomly split into two subsets: a training set (75%) and a test set (25%). The training set was used to develop and optimize the predictive model, while the test set was reserved for evaluating the model's performance on unseen data. To ensure the robustness and reliability of the model, a 10-fold cross-validation approach was employed during the training phase. This approach involves dividing the training data into 10 subsets, iteratively using one subset as validation data and the remaining subsets for training. This process was repeated for multiple configurations to assess the impact of different training and testing ratios on the model's predictive accuracy and generalizability. The final model was selected based on performance metrics, ensuring optimal predictive capacity on the test set. Four ML classifiers were then employed to construct the predictive model. These classifiers were selected based on their suitability for the research problem, the dataset's characteristics, and the research's objective. Each classifier was rigorously trained on the training set and subsequently tested on the testing set to evaluate its performance. Detailed descriptions of these classifiers are provided below:

2.3.1 Gradient boosting (GB)

GB is an ensemble learning classifier that builds models sequentially to minimize prediction errors. It combines the predictive strength of multiple weak learners, typically decision trees, by

optimizing for a loss function [13]. Each tree attempts to correct the errors made by its predecessor, producing a robust and highly accurate model. GB is particularly effective in handling structured data and can adapt to complex patterns. GB excels at capturing subtle relationships and non-linear interactions between customer attributes and churn likelihood for churn prediction, offering high predictive accuracy.

2.3.2 Random forest (RF)

It is an ensemble learning classifier that constructs many decision trees during training and combines their outputs to make a final prediction [14]. It introduces randomness by training each tree on a bootstrap sample of the data and considering only a random subset of features at each split. This reduces the risk of overfitting and enhances generalization. RF is resilient to noisy data and can effectively handle numerical and categorical variables. RF provides reliable and interpretable results in churn prediction, making it suitable for identifying key factors influencing customer retention.

2.3.3 Naive Bayes (NB)

It is a probabilistic classifier that is based on Bayes' theorem, assuming independence between predictors [15]. Despite its simplicity, it is computationally efficient and performs well with high-dimensional data. It is particularly effective for categorical data and works well in scenarios where feature independence is approximately true. For churn prediction, NB offers quick and interpretable insights into customer behavior, especially when the data is well-preprocessed, and the assumption of autonomy holds reasonably well.

2.3.4 Support vector machine (SVM)

It is a supervised learning classifier designed to find an optimal hyperplane that separates data into different classes [16]. It uses a kernel trick to transform non-linear data into a higher-dimensional space where a linear separator is feasible. This classifier is effective in high-dimensional spaces and provides robust performance with well-tuned parameters. For churn prediction, it can model complex decision boundaries between churners and non-churners, particularly in datasets with clear class separability. However, it may require significant preprocessing and parameter tuning for optimal results.

2.4 Evaluation metrics

In ML, evaluating model performance is critical, particularly in domains such as churn prediction, where data imbalance often skews traditional metrics. Accuracy, a widely used measure, is insufficient in imbalanced datasets as it can be dominated by the majority class, leading to misleading results. Precision and recall, which focus on the positive class, offer more nuanced insights by assessing the model's ability to correctly identify churners (precision) and its effectiveness in capturing all actual churners (recall). The F1-score, the harmonic mean of precision and recall, is particularly suited for churn prediction as it balances these metrics,

ensuring neither is disproportionately favored. Additionally, the area under the curve (AUC) has become a standard for evaluating classifier discrimination power, highlighting its ability to differentiate between churners and non-churners across various thresholds. Utilizing these metrics, this study seeks to compare and identify the most effective techniques among SMOTE, ADASYN, and K-means SMOTE to enhance predictive performance in churn modeling.

3. Results

3.1 Performance metrics of ML classifiers with imbalanced data

Table 2 outlines the performance metrics of various ML classifiers when trained on an imbalanced dataset. GB emerges as the most proficient classifier, evidenced by its high AUC of 0.84 and balanced accuracy, F1-score, precision, and recall, indicating strong performance in identifying churn cases. RF follows with a decent AUC of 0.81 and slightly lower metrics, suggesting a good but less effective balance than GB. NB shows a similar AUC (0.82) but lower accuracy and recall, though its precision (0.80) remains high, indicating strength in correctly predicting churn when it does occur. In stark contrast, SVM struggles significantly with the imbalanced data, reflected in its low AUC of 0.33, which indicates poor discrimination between churn and non-churn cases despite a seemingly reasonable accuracy that masks its ineffectiveness in handling class imbalance.

Table 2. Performance of ML classifiers with imbalanced data

Data	Oversampling technique	ML-classifier	AUC	Accuracy	F1-Score	Precision	Recall
Original imbalanced data	Algorithm	GB	**0.84**	**80**	**79**	**79**	**80**
		RF	0.81	78	77	77	78
		NB	0.82	73	75	80	73
		SVM	0.33	75	74	74	75

3.2 SMOTE effects on ML classifier performance

Table 3 presents the impact of applying the SMOTE on ML classifiers, comparing their performance with the original imbalanced dataset. Notably, SMOTE significantly enhances the AUC for GB, RF, and NB, all achieving values of 0.93 or higher compared to their respective AUCs on the imbalanced data. This translates to relative improvements of 12% for GB, 15% for RF, and 14% for NB. Across all other metrics, these classifiers show consistent gains, with RF exhibiting the most substantial accuracy improvement at 9%. NB sees a notable 16% increase in recall, highlighting its enhanced ability to identify churn cases correctly. In contrast, SVM experienced a dramatic 128% improvement in AUC, rising from 0.33 to 0.75 with SMOTE. Yet, it still lags behind the other classifiers in overall metrics, with decreases in accuracy, F1-score, precision, and recall. This suggests that while SMOTE effectively boosts SVM's ability to distinguish between classes, it does not fully remedy its performance issues compared to the other classifiers. Overall,

SMOTE proves to be a valuable technique for improving the predictive capabilities of ML classifiers, particularly enhancing GB's effectiveness in handling previously imbalanced datasets.

Table 3. Performance of ML classifiers using SMOTE

Data	Oversampling technique	ML-classifier	AUC	Accuracy	F1-Score	Precision	Recall
Balanced with SMOTE	Algorithm	GB	**0.94**	**86**	**86**	**86**	**86**
		RF	0.93	85	85	85	85
		NB	0.93	85	85	85	85
		SVM	0.75	69	69	70	69
	Relative improvement in the performance metrics in relation to the original imbalanced data (%)	GB	12%	8%	9%	10%	8%
		RF	15%	9%	10%	10%	9%
		NB	14%	16%	13%	6%	16%
		SVM	128%	-7%	-8%	-5%	-7%

3.3 ADASYN effects on ML classifiers' performance

Table 4 illustrates the impact of ADASYN on the performance of ML classifiers, comparing their results to those on the original imbalanced dataset. As was observed in Table 2, ADASYN significantly boosts the AUC for GB, RF, and NB, with each achieving scores of 0.93 or higher. The classifiers also demonstrate consistent improvements in other metrics, with RF showing a 9% increase in accuracy and NB a 16% rise in recall, indicating better identification of churn cases. While SVM shows a remarkable 130% improvement in AUC, from 0.33 to 0.76, it still trails behind the other classifiers in overall performance, highlighting ongoing difficulties in fully leveraging oversampling benefits. The decline in some metrics for SVM suggests that while ADASYN improves its class distinction, it does not fully resolve its performance issues compared to other classifiers. Overall, ADASYN proves to be an effective technique for enhancing predictive accuracy, particularly improving the performance of GB in the context of imbalanced datasets.

Table 4. Performance of ML classifiers using ADASYN

Data	Oversampling Method	ML-classifier	AUC	Accuracy	F1-Score	Precision	Recall
Balanced with ADASYN	Algorithm	GB	**0.94**	**86**	**86**	**86**	**86**
		RF	0.93	85	85	85	85
		NB	0.93	85	85	85	85
		SVM	0.76	72	71	73	72
	Relative improvement in the performance metrics in relation to original imbalanced data (%)	GB	12%	8%	9%	10%	8%
		RF	15%	9%	10%	11%	9%
		NB	14%	16%	13%	6%	16%
		SVM	130%	-4%	-4%	-2%	-4%

3.4 K-means SMOTE oversampling effects on ML classifier's performance

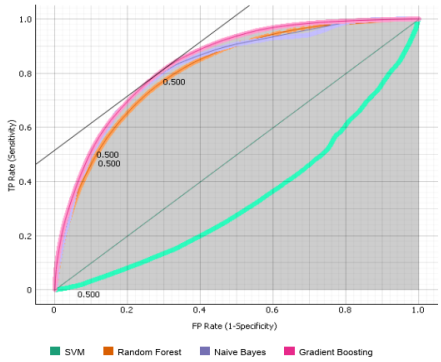
Compared to the original imbalanced data, Table 6 reveals significant improvements in performance metrics after applying K-means SMOTE for oversampling. GB sees a 13% improvement in AUC and consistent enhancements of 9-10% in accuracy, F1 score, precision, and recall, highlighting its strong ability to handle imbalanced data. RF also shows a 15% improvement in AUC, with similar increases in other metrics (9-11%). NB benefits from a 14% AUC improvement and notable gains in recall (16%), though precision improvement is more minor at 7%. However, SVM exhibited the most significant AUC improvement (138%), although its other metrics showed marginal declines. This table highlights K-means SMOTE as the most effective technique for maximizing AUC and maintaining robust performance across classifiers.

Table 5. Performance of ML classifiers using K-means SMOTE

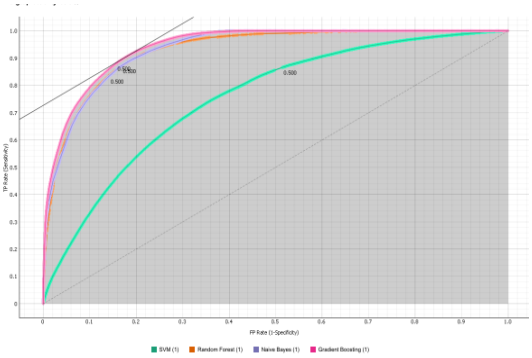
Data	Oversampling Method	ML-classifier	AUC	Accuracy	F1-Score	Precision	Recall
Balanced with K-means SMOTE	Algorithm	GB	0.95	87	87	87	87
		RF	0.93	85	85	85	85
		NB	0.94	85	85	85	85
		SVM	0.79	72	72	72	72
	Improvement in the performance in relation to the original imbalanced data (%)	GB	13%	9%	10%	10%	9%
		RF	15%	9%	10%	11%	9%
		NB	14%	16%	14%	7%	16%
		SVM	138%	-4%	-3%	-3%	-4%

3.5 Performance of ML-based Classifiers with Oversampling Techniques

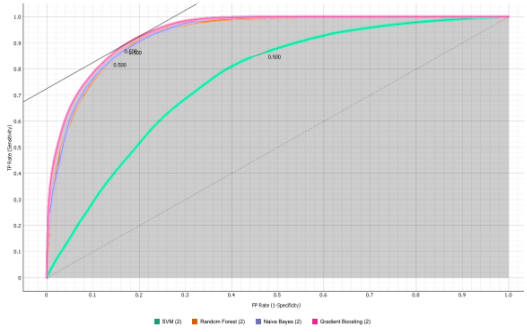
Figure 2 (panels a-d) displays the ROC curves for various ML classifiers applied to the customer churn dataset under four distinct settings: the original dataset, SMOTE, K-means SMOTE, and ADASYN. The classifiers perform differently across these scenarios. On the original dataset, the classifiers generally exhibit subpar performance due to class imbalance, as indicated by lower AUC values. However, there is a noticeable improvement when synthetic sampling techniques are employed. Oversampling techniques enhance the classifiers' ability to distinguish between churn and non-churn customers. Among these, ensemble methods such as RF and GB tend to achieve higher AUC values, showcasing their ability to manage rebalanced datasets effectively. In contrast, simpler classifiers like NB and SVM also benefit, but to a lesser extent. Across all panels, GB consistently stands out as the most effective classifier, particularly excelling in datasets resampled with K-means SMOTE and ADASYN, highlighting its strong capability to tackle class imbalance when combined with sophisticated sampling techniques.



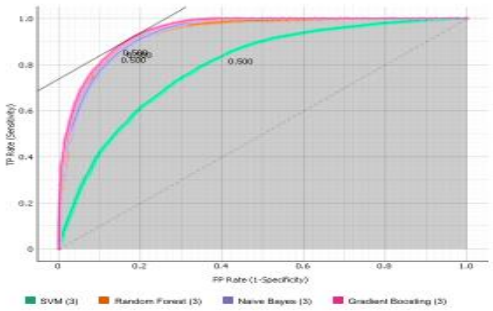
A. Original Dataset



B. Sampled with SMOTE



C. Sampled with K-means SMOTE



D. Sampled with ADASYN

Figure 2. ROC of all classifiers for imbalanced data using SMOTE, K-means SMOTE, and ADASYN.

4. Discussion

Effectively addressing the challenges of imbalanced datasets is crucial for developing reliable ML models in churn prediction. This study provides a fresh perspective on the relationship between oversampling techniques and ML classifiers, shedding light on how data-balancing methods influence classifier performance. The findings highlight the critical role of oversampling techniques in enhancing ML model accuracy and interpretability. Moreover, the conclusions of this study emphasize the importance of continuous evaluation and adaptation of models to reflect changing customer behaviors and market conditions, ensuring that predictive models remain relevant and practical over time.

The study's findings demonstrate that K-means SMOTE is the leading technique for boosting overall model performance, outperforming SMOTE and ADASYN in tackling data imbalance challenges in churn prediction. This conclusion is consistent with previous research, highlighting

K-means SMOTE's superior effectiveness in managing data imbalance compared to other methods. For instance, a study by Liu et al. [5] highlights that K-means SMOTE significantly outperforms traditional methods like SMOTE and ADASYN in managing class imbalance, leading to improved prediction accuracy in churn prediction. Similarly, research by Douzas et al. [17] demonstrated that K-means SMOTE excels across various testing scenarios, confirming its effectiveness in enhancing classifier performance for churn datasets. Moreover, Azal et al. [18] found that K-means SMOTE consistently delivered superior results across multiple imbalanced churn prediction datasets. Lastly, a study by Ong et al. ([19] reaffirmed K-means SMOTE's distinct advantage over other oversampling methods, solidifying its status as a leading technique for addressing data imbalance in churn prediction tasks.

This study's findings further reveal that GB consistently stands out as the most effective classifier compared to RF and NB for predicting churn. Recent research consistently affirms that GB is a leading classifier for churn prediction. AlShourbaji et al. [6] conducted a thorough analysis showing that GB surpasses ML classifiers like RF and NB in predictive accuracy thanks to its adept handling of complex data patterns. Zhenkun et al. [20] further highlighted its efficiency in optimizing class weights and hyperparameters, enhancing churn prediction results. Likewise, Ogbonna et al. [21] found that GB achieves superior precision and recall rates, making it the favored choice for churn tasks. Mouli et al. [22] also concluded that GB classifiers excel in accuracy and reliability over other classification models. Together, these studies confirm the superiority of GB, supporting our findings.

5. Conclusion

This study emphasizes the critical role of addressing data imbalance in churn prediction for subscription-based businesses. By exploring advanced oversampling techniques like K-means SMOTE, SMOTE, and ADASYN, the study demonstrated how these techniques can effectively enhance the predictive accuracy of ML classifiers that identify potential churners. Integrating these techniques with the ML classifiers provides a robust framework for tackling the challenges of imbalanced datasets. This approach improves the models' ability to detect churners and ensures a fairer allocation of resources in predictive analytics. The insights gained from this research hold valuable implications for the broader ML field, particularly in applications beyond churn prediction. By focusing on the interaction between oversampling techniques and ML classifiers, this study contributes to optimizing predictive models in various domains. Future research can build upon this foundation by exploring additional techniques and evaluating their applicability across different datasets and industries, further enhancing the capabilities of ML in managing class imbalance.

Acknowledgements: The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FFR-2025-1635-07".

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] A. Sikri, R. Jameel, S.M. Idrees, H. Kaur, Enhancing Customer Retention in Telecom Industry with Machine Learning Driven Churn Prediction, Sci. Rep. 14 (2024), 13097. https://doi.org/10.1038/s41598-024-63750-0.

[2] A. Rodan, A. Fayyoumi, H. Faris, J. Alsakran, O. Al-Kadi, Negative Correlation Learning for Customer Churn Prediction: A Comparison Study, Sci. World J. 2015 (2015), 473283. https://doi.org/10.1155/2015/473283.

[3] A. H M, B. T, S. Tanisha, S. B, C.C. Shanuja, Customer Churn Prediction Using Synthetic Minority Oversampling Technique, in: 2023 4th International Conference on Communication, Computing and Industry 6.0 (C216), IEEE, 2023, pp. 01-05. https://doi.org/10.1109/C2I659362.2023.10430989.

[4] S.J. Haddadi, A. Farshidvard, F.D.S. Silva, J.C. dos Reis, M. da Silva Reis, Customer Churn Prediction in Imbalanced Datasets with Resampling Methods: A Comparative Study, Expert Syst. Appl. 246 (2024), 123086. https://doi.org/10.1016/j.eswa.2023.123086.

[5] X. Liu, G. Xia, X. Zhang, W. Ma, C. Yu, Customer Churn Prediction Model Based on Hybrid Neural Networks, Sci. Rep. 14 (2024), 30707. https://doi.org/10.1038/s41598-024-79603-9.

[6] I. AlShourbaji, N. Helian, Y. Sun, A.G. Hussien, L. Abualigah, et al., An Efficient Churn Prediction Model Using Gradient Boosting Machine and Metaheuristic Optimization, Sci. Rep. 13 (2023), 14441. https://doi.org/10.1038/s41598-023-41093-6.

[7] I. Ullah, B. Raza, A.K. Malik, M. Imran, S.U. Islam, et al., A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector, IEEE Access 7 (2019), 60134-60149. https://doi.org/10.1109/access.2019.2914999.

[8] D.T. Barus, R. Elfarizy, F. Masri, P.H. Gunawan, Parallel Programming of Churn Prediction Using Gaussian Naïve Bayes, in: 2020 8th International Conference on Information and Communication Technology (ICoICT), IEEE, 2020, pp. 1-4. https://doi.org/10.1109/ICoICT49345.2020.9166319.

[9] Telco Customer Churn Dataset, Telco Customer Churn, https://www.kaggle.com/blastchar/telco-customer-churn, Accessed Oct. 2, 2024.

[10] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-Sampling Technique, J. Artif. Intell. Res. 16 (2002), 321-357. https://doi.org/10.1613/jair.953.

[11] M. Mujahid, E. Kına, F. Rustam, M.G. Villar, E.S. Alvarado, et al., Data Oversampling and Imbalanced Datasets: An Investigation of Performance for Machine Learning and Feature Engineering, J. Big Data 11 (2024), 87. https://doi.org/10.1186/s40537-024-00943-4.

[12] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, 2008, pp. 1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969.

[13] J.H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, Ann. Stat. 29 (2001), 1189–1232. https://doi.org/10.1214/aos/1013203451.

[14] L. Breiman, Random Forests, Mach. Learn. 45 (2001), 5-32. https://doi.org/10.1023/a:1010933404324.

[15] P. Domingos, M. Pazzani, On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss, Mach. Learn. 29 (1997), 103-130. https://doi.org/10.1023/a:1007413511361.

[16] B. Schölkopf, A.J. Smola, Learning with Kernels, The MIT Press, 2001. https://doi.org/https://doi.org/10.7551/mitpress/4175.001.0001.

[17] G. Douzas, F. Bacao, F. Last, Improving Imbalanced Learning Through a Heuristic Oversampling Method Based on K-Means and SMOTE, Inf. Sci. 465 (2018), 1-20. https://doi.org/10.1016/j.ins.2018.06.056.

[18] A.A. Khan, O. Chaudhari, R. Chandra, A Review of Ensemble Learning and Data Augmentation Models for Class Imbalanced Problems: Combination, Implementation and Evaluation, Expert Syst. Appl. 244 (2024), 122778. https://doi.org/10.1016/j.eswa.2023.122778.

[19] J. Ong, G. Tong, K. Khor, S. Haw, Enhancing Customer Churn Prediction with Resampling: A Comparative Study, TEM J. 13 (2024), 1927-1936. https://doi.org/10.18421/tem133-20.

[20] Z. Liu, P. Jiang, K.W. De Bock, J. Wang, L. Zhang, et al., Extreme Gradient Boosting Trees with Efficient Bayesian Optimization for Profit-Driven Customer Churn Prediction, Technol. Forecast. Soc. Chang. 198 (2024), 122945. https://doi.org/10.1016/j.techfore.2023.122945.

[21] O.J. Ogbonna, G.I. Aimufua, M.U. Abdullahi, S. Abubakar, Churn Prediction in Telecommunication Industry: A Comparative Analysis of Boosting Algorithms, Dutse J. Pure Appl. Sci. 10 (2024), 331-349. https://doi.org/10.4314/dujopas.v10i1b.33.

[22] K.C. Mouli, C.V. Raghavendran, V.Y. Bharadwaj, G.Y. Vybhavi, C. Sravani, et al., An Analysis on Classification Models for Customer Churn Prediction, Cogent Eng. 11 (2024), 2378877. https://doi.org/10.1080/23311916.2024.2378877.