sciendo

Performance Analysis of Data Balancing Methods for Churn Prediction

Yanka ALEKSANDROVA*
University of Economics-Varna, Varna, Bulgaria
**Corresponding author, yalexandrova@ue-varna.bg*

Desislava KOLEVA
University of Economics-Varna, Varna, Bulgaria
desi_koleva@ue-varna.bg

Abstract. *This study evaluates the influence of various data balancing techniques on the performance of machine learning models for churn prediction across multiple imbalanced datasets. The proposed approach consists of data preparation, application of data balancing techniques on the training data, model training with hyperparameter optimization using genetic algorithms and comparative performance evaluation of the trained models. Six balancing techniques are evaluated —Random Undersampling, Random Oversampling, SMOTE, SMOTEENN, KMeansSMOTE, and ADASYN. The machine learning algorithms chosen are ensembles, such as Random Forest, Gradient Boosting Machines and XGBoost. Results indicate that XGBoost consistently outperforms other models, particularly when used in combination with SMOTE and SMOTEENN, achieving the highest sensitivity, F1 score and overall performance. Random Forest also reveals excellent predictive capabilities, especially with regard to correctly classifying loyal customers. SMOTE and SMOTEENN, particularly in combination with XGBoost and GBM, stand out as the most effective data balancing techniques, significantly improving model sensitivity. SMOTE performs particularly well when used with XGBoost and GBM, while SMOTEENN improves Random Forest's ability to detect churners. The findings highlight the importance of selecting the appropriate algorithm and balancing technique based on dataset characteristics, business requirements and objectives of customer retention strategies.*

Keywords: churn prediction, machine learning, data balancing techniques, SMOTE, SMOTEENN, ensemble machine learning

Introduction

Proper customer relationship management is a key factor in forming high financial results, increased market share, and reducing costs in companies. Important indicators, to track the current state of the company and to make the right business decisions, are the number of new customers, number of retained customers and churn customers. According to Matthew-El et al. (2024) it often takes five to 20 times the number of resources to obtain a new customer than to retain an existing one. This focuses a special attention of the companies to the number of churn customers and on developing effective strategies to build loyalty and prevent customer churn.

Successful business strategies heavily depend on churn prediction because it helps organizations understand which customer relationships need proactive management (Sun, et al., 2023). The analysis of historical data, combined with customer behavior patterns, enables organizations to detect indicators of customer deactivation (Wang, et al., 2024). This allows the companies to proactively intervene early with implementing retention efforts, ensuring that valuable customers are not lost before addressing the underlying problems. Organizations, adopting

DOI: 10.2478/picbe-2025-0074

this proactive approach, maintain their revenue streams and reduce the substantial expenses involved in acquiring new customers.

Customer churn can be generally measured as the number of customers that stop buying products (Thamizhselvi & al., 2024) or rate at which customers switch to competitors (Basha & el., 2024). Predicting churn is aiming to sustain growth and profitability of the business. The benefits of predicting churning customers are related to receiving information about customer behavior and preferences that will help to create suitable retention strategies, improve product development, marketing, customer service and make informed business decisions about resource allocation for maximum impact. Companies need to adopt a proactive strategy, addressing potential issues before they lead to customer loss. It is crucial for the churn prediction to be made at the earliest possible moment, before it happened and to identify the possible reasons for it (Bhattacharjee et al., 2023). Because of the dynamics of the business, this task has become a challenge for every company. The customer churn is directly related to customer retention, because the churn will increase retention.

Churn prediction has been traditionally implemented using statistical methods like logistic regression (Agarwal, 2024) and survival analysis (Denayer, 2024; Sharma, 2024) because of their interpretability and simple implementation. The models provide clear understanding of how each variable affects churn probability. The simplicity of these models, however, limits their effectiveness when dealing with non-linear variable relationships with multiple interdependent elements. The assumptions of statistical models restrict their predictive capability and as a result, they deliver inferior results than advanced methods for churn prediction.

Machine learning (ML) methods have recently become popular, and they are preferable due to the capability of overcoming the weaknesses of statistical and scoring approaches like Recency Frequency Monetary (RFM) analysis. The main advantage of using ML technology is the ability to handle large volumes of data, identifying non-linearity and complex relationships and uncover patterns, that might not be explicit (Ahmed et al., 2024). All these characteristics lead to higher predictive accuracy of ML models compared to statistical methods (Sun, 2023; Wang, 2024; Ren, 2025). Another advantage of machine learning algorithms is the ability to improve their efficiency and accuracy over time, because of the automated feature continuous learning.

At the same time, there are some challenges of building machine learning predictive models (Shatishkumar, 2023). ML technology needs a large amount of data to create successful training models. A less amount of training data will produce inaccurate or too biased predictions (Atay & Turani, 2024; Wang, 2024; Demirbaga et al., 2024). In some cases, companies do not have enough data, and/or do not use an approach to systematically collect and analyze data.

Another challenge is disbalancing datasets which makes it difficult to train models to correctly classify the minority class. This problem is particularly evident in the area of customer churn prediction. The percentage of churners can vary across different industries, although we can assume that churn events are far less frequent than non-churn ones. Different researchers have studied datasets for churn prediction and reported churn percentages ranging from below 10% (Burez & Van den Poel, 2009) to 26% (Wagh et al., 2024). This imbalance causes difficulty of accurately identifying churners, as the predictive models must learn to detect rare events amid a predominance of non-churn data. The development of valuable churn prediction models for business retention requires successful resolution of the problem with imbalanced datasets.

Choosing an appropriate machine learning algorithm is very crucial to building accurate models for churn prediction (Lalwani et al., 2021; Loukili et al., 2022). In most cases this requires different types of algorithms to be applied on data, hyperparameter optimization and evaluation

performance analysis in order to choose the best predictive model (Poudel et al., 2024; Haddadi et al., 2024). The purpose of this conference report is to compare and analyze different approaches for solving the problem with imbalanced datasets as part of the whole process for training, optimization and evaluation of machine learning models for churn prediction.

Literature review

Customer churn prediction models play an important role in organizations trying to retain customers in saturated and rapidly changing markets (Sun, et al., 2023).

The customer churn problem is defined as a classification problem in which it is necessary to make a prediction whether the customer will switch to competitors, stop buying products, cancel services or stop interacting with the company. The purpose of the classifier is to correctly identify leaving clients (churners) as positive class 1 and non-churners as negative class 0, thus defining a binary classification problem (Ljubičić et al., 2023; Lalwani et al., 2021; Loukili et al., 2022).

Machine learning algorithms

Researchers report experimental results of using different types of ML algorithms for solving customer churn problems. The most popular single classifiers are:

- *Logistic Regression (LR)* - a statistical method that predicts the probability of a binary outcome (e.g. churn or non-churn) based on one or more predictor variables. Its main advantages are simplicity and interpretability (Loukili et al., 2022; Awodele et al., 2020; Wang et al., 2024; Pratheebha et al., 2024).
- *Decision Trees (DT)* - a tree-based algorithm which is easy to be interpreted and that can handle both numerical and categorical data (Awodele et al., 2020; Lalwani, et al., 2021; Wagh et al., 2024; Shumaly et al., 2020).
- *Naive Bayes (NB*) - a probabilistic classifier based on Bayes' theorem that assumes independence between the features. This algorithm is popular because of its simplicity and efficiency, especially with large datasets (Awodele et al., 2020; Lalwani et al., 2021).
- *Support Vector Machines (SVM)* - a supervised learning algorithm that analyzes data for classification and regression analysis by finding the hyperplane that best divides a dataset into classes. Effective in high-dimensional spaces and used for its robustness (Loukili et al., 2022; Kun et al., 2024; Shumaly et al., 2020).
- *K-Nearest Neighbors (kNN)* - a non-parametric method used for classification and regression, where the output is a class membership, or a property value based on the k closest training examples in the feature space. It is simple and effective for small datasets (Loukili et al., 2022; Prabadevi et al., 2023; Lalwani et al., 2021; Wagh et al., 2024).

Another important group of algorithms includes *ensemble methods* which combine multiple models to improve the overall performance of the classifier. Common ensemble methods, applied for churn prediction, include:

- *Random Forest (RF)* - combines multiple decision trees to improve accuracy and control overfitting (Loukili et al., 2022; Shumaly et al., 2020; Wang et al., 2024).
- *Gradient Boosting Machines (GBM)* - builds models sequentially, each new model correcting errors made by the previous ones (Prabadevi et al., 2023).
- *AdaBoost* - focuses on the instances that previous classifiers misclassified, adjusting their weights accordingly (Lalwani et al., 2021; Poudel et al., 2024).
- *XGBoost* - optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable (Lalwani et al., 2021; Wang et al., 2024).

- CatBoost - handles categorical features automatically and is robust to overfitting (Haddadi et al., 2024).

In addition, there are many publications documenting the successful implementation of *Neural Networks (NN)* for churn prediction. Various NN-based algorithms have been explored, like Multilayer Perceptron, Convolutional Neural Networks, Back Propagation Neural Network (Wang et al., 2024; Awodele et al., 2020; Kun, et al., 2024) among others.

The accuracy of the churn prediction models reported in scientific publications varies depending on the chosen dataset and applied algorithms. Prabadevi et al. (2023) documented the implementation of Stochastic Gradient Booster, Random Forest, Logistics Regression, and k-Nearest neighbors (kNN) methods with accuracy of 0.839, 0.826, 0.829 and 0.781 respectively Other authors compare the performance of classification algorithms such as Random Forest (RF), kNN and Decision Tree Classifier (DT) and achieved accuracy of 0.99. Wagh et al. (2024) reports the predominance of the Fully Connected Layer Convolutional Neural Network - Long Short-Term Memory (FCLCNN-LSTM) based on Deep Learning. It has been compared with LR, RF, SVM, XGBoost, CNN, LSTM (Long Short-Term Memory). The team has tested three datasets in different industries: insurance, telecommunications and banking. FCLCNN-LSTM has shown the highest accuracy as 0.9415 (insurance), 0.8121 (telecom industry), 0.8651 (banking). Poudel et al (2024) reported the best performance of accuracy metric using GBM (0.81) in comparison of RF (0.80), XGB (0.80), LR (0.79), AdaBoost (0.79), SVC (Support Vector Classifier) (0.78), NN (Neural networks) (0.74). Lalwani, et al. (2021) found that AdaBoost and XGboost classifier give the highest accuracy of 0.817 and 0.808 respectively. The highest AUC score of 0.84 in the same research is achieved by both AdaBoost and XGBoost classifiers which outperform the others.

Data balancing methods

Imbalanced training data, where one class significantly outnumbers others, can significantly affect the performance of machine learning models (Haddadi et al., 2024; Poudel et al., 2024). Models trained on imbalanced datasets are biased towards the majority class, leading to poor performance in predicting the minority class (Prabadevi, 2024). This bias results in high accuracy for the majority class but low accuracy for the minority class, which is often critical in applications like fraud detection or medical diagnosis. Another challenge is overfitting, as imbalanced data increases the risk of overfitting, where the model memorizes the majority class samples instead of learning the underlying patterns (Sánchez-Gutiérrez & González-Pérez, 2023). This overfitting reduces the model's ability to generalize when applied on new, unseen data.

To address the issue of imbalanced datasets in machine learning, several data balancing methods can be used. These methods aim to improve the performance of machine learning models by ensuring that the minority class is adequately represented. Some suitable methods are:

Oversampling methods. In this group of methods, the number of cases from the less represented class is increased to achieve the desired ratio. Several methods can be distinguished like:

- *Random Over Sampler*. The method randomly duplicates cases from the minority class until the desired class proportions are achieved. It is a simple technique, but it can lead to overfitting since there is no new information added to the dataset.
- *SMOTE (Synthetic Minority Over-Sampling Technique)*. SMOTE generates new synthetic cases by interpolating existing minority class representatives. The new points are created by aggregating information about the k-nearest neighbors of cases from the minority class.

- *SMOTEENN (SMOTE with Edited Nearest Neighbors).* This method adjusts SMOTE by using ENN to remove noisy data and cases which are close to the majority class, thus improving class separation.
- *ADASYN (Adaptive Synthetic Sampling)* is another extension of SMOTE which generates synthetic samples based on the minority class density distribution. As a result, more synthetic cases are created in spaces where there are far less minority cases.
- *KMeansSMOTE.* This method uses K-Means clustering algorithm before the implementation of SMOTE. New samples are generated within each previously identified cluster.

Undersampling methods. As the name implies, fewer cases from the majority class are used for training to achieve the desired class ratio (usually 50:50). Such method is Random Over-Sampling Example (ROSE) generating random new cases to achieve a balanced data set.

Mixed techniques. These methods are aimed at selecting fewer cases from the majority class and more cases from the minority class to achieve the desired ratio.

In their research of customer churn prediction with ML, Wagh et al. (2024) have published the results of evaluation of training model on an imbalanced dataset and after balancing it. Experimental analysis using a decision tree classifier obtained accuracy of the model 0.78. After using SMOTEENN, the dataset is balanced, and analysis using decision tree classifier obtains accuracy of the model 0.9385. Within the same experimental analysis random forest algorithms achieve accuracy of the model 0.9891 before and 0.9909 after balancing with SMOTEENN. Haddadi at al. (2024) reports very good performance of SMOTE and ADASYN resampling techniques. They have compared AUC coefficients before and after using three techniques for balancing on three datasets. The techniques used are SMOTE, ADASYN and two-phase approach, that incorporates elements of both data-level under sampling and ensemble-based methods. The results from telecommunication dataset are AUC of 0.84 for the two-phase approach, SMOTE and ADASYN using Random Forest algorithm against 0.69 of AUC no balancing data. The same result (AUC=0.84) is achieved using SMOTE, ADASYN and Cat Boost algorithm. For comparison the value of AUC on balanced data was 0.7. The second-best balancing techniques reported by the same team is AUC of 0.83 for models using SMOTE and ADASYN in combination with XGBoost and GBM, compared to AUC of 0.7 AUC no balanced data.

Using balancing techniques in online retail dataset (Haddadi et al., 2024) generates the best AUC, equal to 0.97, for models using SMOTE and ADASYN with XGBoost , second-best result is AUC=0.96 for the combinations of SMOTE and ADASYN with Random forest (AUC=0.88 with imbalanced dataset) and models using SMOTE and ADASIN with CatBoost (0.87 on imblanaced dataset).The evaluation results in the same experiments on the banking dataset in terms of AUC show the best performance (0.86) of SMOTE with XGBoost, SMOTE with Random Forest and SMOTE with CatBoost against respectively (0.70, 0.68, 0.71 with imbalanced data).

Methodology

We propose an approach that will systematically assess the influence of various balancing techniques on the performance of machine learning models trained on multiple imbalanced datasets for churn prediction. The approach consists of multiple stages, represented in fig.1. The first stage is data preparation in which we perform exploratory data analysis to examine the dataset structure, identify missing values, outliers, wrong formats, etc. After addressing data quality issues, we perform standard scaling on the numerical variables and one hot encoding on categories. The whole dataset is split into training and testing datasets.

The current study will consider six balancing methods – Random Undersampling, Random Oversampling, SMOTE, SMOTEENN, KMeansSMOTE and ADASYN. Each technique offers a unique approach to handling minority class representation, and by applying these methods, we can investigate their influence on the machine learning models performance. For comparison purposes we also train respective models on the original imbalanced dataset.

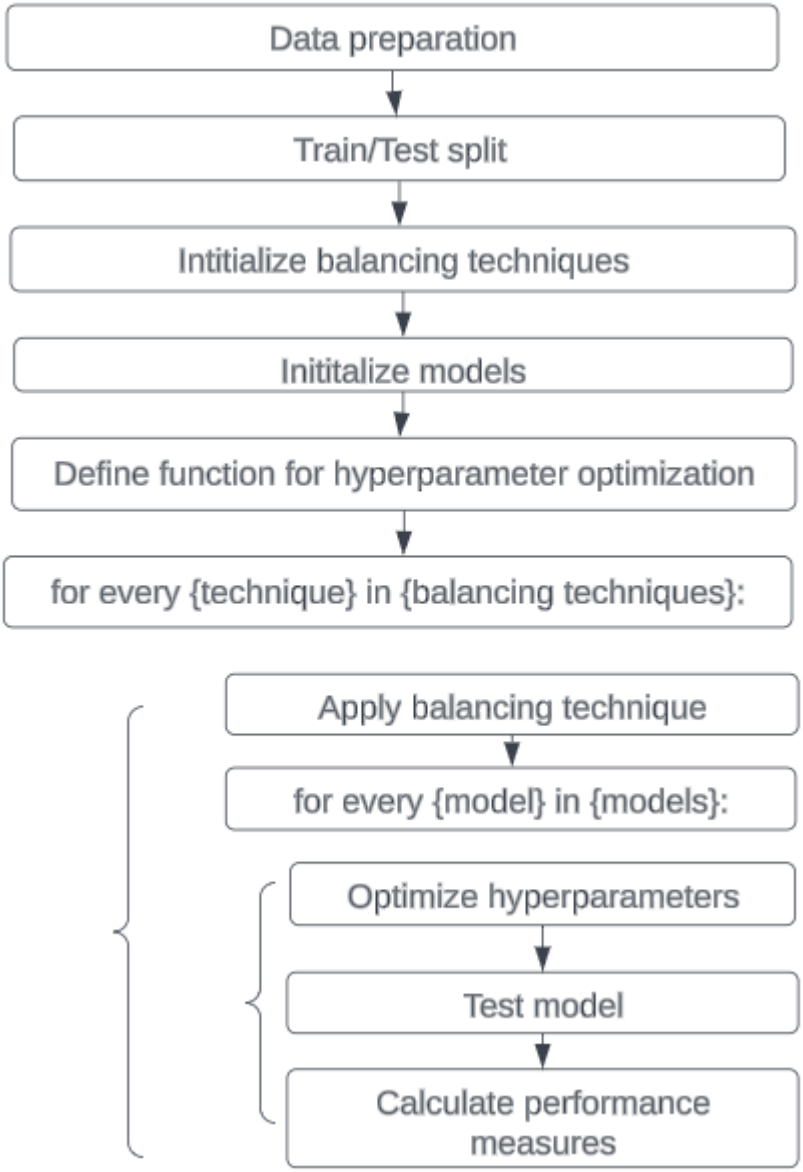

Figure 1. Proposed approach for comparing balancing techniques
Source: Authors' own research.

The effectiveness of machine learning models depends heavily on hyperparameter optimization because it determines model performance. Missing or poor optimization of models results in performance problems which create inaccurate predictions along with reduced capability to generalize. The various optimization methods aim at selecting the best parameters to train a model that works best for its targeted application. Two primary basic methods in hyperparameter

optimization are grid search where predefined parameter combinations are evaluated, while random search randomly selects parameter configurations. Bayesian optimization applies probabilistic techniques to model how hyperparameters perform and select optimal candidates through analysis of previous evaluation results (Frazier, 2018).

Another robust alternative for hyperparameter optimization is genetic algorithms. These algorithms generate a population of candidate parameters, and then select the best-performing ones, recombine them through crossover and mutation to produce improved generations. This evolutionary approach enables a comprehensive exploration of the hyperparameter space, thus overcoming the limitations of the traditional methods like grid or random search. One of the biggest advantages of genetic algorithms is their capability to avoid local sub-optimal solutions which makes them very effective in handling complex parameter spaces (Yokoyama et al., 2023; Gour et al., 2024). As a result, genetic algorithms can be especially effective in optimizing machine learning models to achieve high performance.

Considering the advantages of genetic algorithms, we chose them as approach hyperparameter optimization and implemented them within the DEAP framework (De Rainville et al., 2012). We defined function which is designed to maximize the objective measure – AUC from cross validation. AUC is selected as a complex measure for the overall model's performance. The hyperparameters chosen to be optimized are max_depth, n_estimators and learning rate.

To train the models, we chose representatives of the ensemble decision trees-bases algorithms, namely Random Forest (RF), XGBoost and Gradient Boosting Machine (GBM). Evident from the literature review, these algorithms have been consistently distinguished as one of the best performing for churn prediction. Every algorithm is trained with optimal hyperparameters on different training sets produced by applying the chosen balancing techniques. All fitted models are then applied to the one and same test dataset to ensure objective comparison. A set of evaluation metrics, result from models' testing, are calculated and analyzed. This set of metrics include AUC, balanced accuracy, F1 score, sensitivity and specificity, which will allow for a comprehensive comparison and results evaluation. To combine all measures, we also calculated „Overall Score" metric as averaged value of AUC, balanced accuracy, F1, sensitivity and specificity.

Limitations of the current research are related to the chosen datasets, algorithms and balancing techniques. We have chosen datasets enabling churn prediction in different industries. Although the chosen datasets cover various industries, we cannot assume that this dataset collection is representative of all possible economic areas or industries. Nevertheless, by using more than one diverse dataset we try to ensure a more comprehensive comparison of the chosen balancing techniques. Based on the literature review and the results presented from other relevant research papers discussed in this section, we chose the best performing ensemble algorithms like XGBoost, Random Forest and GBM. Hence, the results can be interpreted with regard to the selected algorithms, datasets and balancing techniques. The proposed approach (see fig. 1), however, can be implemented with different algorithms, datasets and balancing techniques and results from the experiments can be further analyzed to compare the performance of a broader set of techniques and algorithms.

All practical experiments were implemented in Jupyter Lab environment using Python packages such as scikit learn, DEAP, xgboost and imblearn among others.

Results and discussions

To ensure a more comprehensive comparison of the chosen balancing techniques, we used different datasets for churn predictions. A summary of datasets is provided in Table 1. As evident from the

table all datasets are heavily imbalanced with proportion of churners varying from 16.70% to 26.50%. All datasets are downloaded from Kaggle platform and cover different industries like telecom services, e-commerce, bank and credit card services.

Table 1. Description of the used datasets

Dataset	Industry	Number of cases	Categorical factors	Numerical factors	Churn rate
E-commerce[1]	e-commerce	9288	5	18	16,70%
Telco[2]	telecom services	7043	11	8	26,50%
Bank[3]	bank services	10000	3	11	20,30%
Churn[4]	credit card services	28382	12	3	18,50%

Source: Authors' own research.

According to the proposed approach (see fig.1) we trained models using 3 ensemble algorithms (Random Forest, GBM, XGBoost) with 7 balancing techniques (Random Under Sampling, Random Over Sampling, SMOTE, SMOTEENN, KMeans SMOTE, ADASYN, None), each combination of algorithm and technique was trained on every dataset. During the training hyperparameter optimization using DEAP was performed on every model. As a result, we collected results from the performance of 84 models. The results for the top five best performing models according to the different metrics and across datasets are shown in Table 2.

Results reveal clear performance trends of different machine learning algorithms and balancing techniques when combined for predicting customer churn across selected datasets. XGBoost proves to be the top-performing algorithm since it is constantly achieving the highest AUC, F1 score and overall score. It is particularly powerful when used with SMOTE as a balancing method, because SMOTE enhances the ability to deal with class imbalances by creating synthetic samples. This combination performs exceptionally well across different datasets and proves to be an optimal solution for churn prediction.

Random Forest also demonstrates excellent performance, especially in specificity, which means that the trained with Random Forest models can minimize false positive errors. It is recommended that the aim is to reduce the incorrect classification of loyal customers (non-churners). Random Forest models when used with SMOTEENN and ADASYN reveal one of the highest values of sensitivity, especially with Telco and Bank datasets.

Regarding the balancing techniques, SMOTE and SMOTEENN can be distinguished as the most effective ones. SMOTE provides excellent results, especially when used with XGBoost and GBM. SMOTEENN with XGBoost combination stands out with its sensitivity showing excellent capability to correctly classify customers, who are likely to churn. SMOTEENN proves to be also very effective when used as a balancing technique with Random Forest, ensuring one of the top values of sensitivity.

When we analyze the performance of different models across the four datasets, we can conclude that different datasets benefit from different combinations of algorithm and balancing

[1] https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction [Last access on 14.0.2025]
[2] https://www.kaggle.com/datasets/blastchar/telco-customer-churn [Last access on 14.03.2025]
[3] https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling [Last access on 14.03.2025]
[4] https://www.kaggle.com/datasets/shubh0799/churn-modelling [Last access on 14.0.2025]

techniques. These findings reveal the importance of the process of selecting the right algorithm and balancing techniques according to the specific dataset characteristics.

Table 2. Top 5 models by metrics and datasets

Bank dataset		Churn dataset		E-commerce dataset		Telco dataset	
Top 5 models by AUC							
XGB RandUnder	0,8504	XGB RandUnder	0,8535	GBM None	1,00000	XGB RandUnder	0,8456
XGB RandOver	0,8503	XGB RandOver	0,8514	GBM RandOver	1,00000	XGB None	0,8452
XGB SMOTEENN	0,8500	RF RandUnder	0,8511	XGB None	0,99998	GBM RandUnder	0,8451
GBM RandUnder	0,8500	RF RandOver	0,8504	GBM ADASYN	0,99998	GBM None	0,8443
RF RandOver	0,8487	GBM RandUnder	0,8470	XGB ADASYN	0,99998	RF None	0,8440
Top 5 models by F1							
XGB SMOTE	0,6073	RF RandOver	0,6149	GBM None	1,00000	RF SMOTE	0,6461
GBM SMOTE	0,6033	GBM RandOver	0,6051	GBM ADASYN	0,99515	GBM RandOver	0,6359
GBM ADASYN	0,5911	XGB RandOver	0,6050	RF RandOver	0,99515	RF ADASYN	0,6355
XGB ADASYN	0,5868	XGB RandUnder	0,5938	GBM SMOTE	0,99515	XGB RandUnder	0,6347
RF RandOver	0,5863	GBM RandUnder	0,5931	RF KMSMOTE	0,99515	RF RandOver	0,6337
Top 5 models by Sensitivity							
XGB SMOTEENN	0,7764	XGB SMOTEENN	0,7570	GBM None	1,00000	RF SMOTEENN	0,8538
RF SMOTEENN	0,7469	GBM SMOTEENN	0,7493	RF SMOTE	1,00000	RF ADASYN	0,8360
RF RandUnder	0,7346	RF RandUnder	0,7473	GBM RandUnder	1,00000	XGB ADASYN	0,8164
XGB RandOver	0,7273	RF SMOTEENN	0,7337	RF RandUnder	1,00000	GBM SMOTEENN	0,8111
GBM RandUnder	0,7273	XGB RandUnder	0,7308	GBM RandOver	1,00000	GBM RandOver	0,8111
Top 5 models by Specificity							
RF None	0,9755	RF KMSMOTE	0,9503	GBM None	1,00000	RF None	0,9090
XGB None	0,9617	GBM KMSMOTE	0,9455	XGB None	1,00000	GBM None	0,9064
GBM None	0,9592	XGB KMSMOTE	0,9446	XGB KMSMOTE	1,00000	XGB None	0,9038
GBM KMSMOTE	0,9159	RF RandOver	0,9044	GBM KMSMOTE	1,00000	XGB KMSMOTE	0,8586
GBM SMOTE	0,8820	GBM RandOver	0,8887	XGB RandOver	1,00000	GBM KMSMOTE	0,8560
Top 5 models by Overall Score							
XGB SMOTE	0,7518	XGB RandOver	0,7620	GBM None	1,00000	RF SMOTE	0,7618
RF RandOver	0,7434	XGB RandUnder	0,7589	GBM RandOver	0,99845	RF ADASYN	0,7596
GBM SMOTE	0,7428	RF RandUnder	0,7571	GBM ADASYN	0,99845	GBM RandOver	0,7588
XGB ADASYN	0,7428	GBM RandUnder	0,7559	RF RandOver	0,99845	XGB RandUnder	0,7570
GBM ADASYN	0,7425	RF RandOver	0,7536	GBM SMOTE	0,99845	GBM RandUnder	0,7559

Source: Authors' own research.

When we analyze the performance of different models across the four datasets, we can conclude that different datasets benefit from different combinations of algorithm and balancing

techniques. These findings reveal the importance of the process of selecting the right algorithm and balancing techniques according to the specific dataset characteristics.

Overall, results from models' evaluation suggest that XGBoost is the best overall algorithm, particularly when paired with SMOTE and SMOTEENN for churn prediction. The choice of balancing technique plays an important role in enhancing model performance.

Numerous recent research studies support our findings which demonstrate the necessity of combining data balancing techniques with advanced ML algorithms and hyperparameter optimization to improve the churn prediction in various domains. One of our main findings is that using ensemble models like XGBoost, GBM and Random Forest, in combination with SMOTE and SMOTEENN leads to better performance in AUC, sensitivity and F1-score. This finding is also supported by Pratheebha et al. (2024), who developed models integrating SMOTEENN, Random Forest and K-Means clustering.

Haddadi et al. (2024) evaluated resampling methods and found that SMOTE and ADASYN, when applied with ensemble classifiers like Random Forest and CatBoost, significantly improved AUC values from 0.69 (imbalanced data) to 0.97 on balanced datasets. This is closely aligned with our results, where XGBoost and GBM models in combination with SMOTE and SMOTEENN achieved one of the best AUC values across all datasets. These support the conclusion that oversampling techniques are one of the most effective approaches for better prediction on the minority class.

The advantages of hybrid balancing techniques like SMOTEENN and SMOTETomek are also acknowledged by Alhakim et al. (2024). A significant improvement SMOTE and SMOTEENN for churn prediction is reported in the study of Harshimi et al. (2024) who used these oversampling techniques with combination of SVM and Ako et al. (2024) who developed models using combination with Random Forest and XGBoost machine learning algorithms.

The role of hyperparameter optimization, a key stage in our proposed approach, is emphasized by Sarkar et al. (2024) who demonstrated that tuned models applied on SMOTEENN balancing training dataset significantly outperformed models without hyperparameter optimization. Strong evidence confirming our findings about the effectiveness SMOTE and ADASYN is provided also by Imani et al. (2024), achieving F1-score of 89% and AUC of 95% with ensemble models.

Based on the results achieved we can summarize our findings in different dimensions. The balancing techniques improve significantly model performance. Data balancing techniques, especially SMOTE and SMOTEENN, largely enhanced the overall accuracy and ability to correctly classify the minority class (churners). Among the evaluated algorithms, XGBoost consistently showed one of the best results with regard to different metrics like AUC, F1-score and overall evaluation score, especially when implemented on dataset balanced with SMOTE and SMOTEENN. The use of genetic algorithms for hyperparameter optimization significantly improved model performance in comparison to models trained with default parameters. At the same time, it's worth mentioning that no single model or data balancing technique dominates across all datasets as evident from the results in table 2. This affirms the importance of dataset-specific model and balancing technique selection.

The proposed approach proved to deliver satisfactory results across datasets from different industries (telecom, banking, credit card services, e-commerce), making it a practical solution when dealing with churn prediction on imbalanced datasets. A substantial theoretical contribution is the development of the framework of the approach for systematic selection, optimization and evaluation of models' performance for churn prediction using different data balancing techniques

and algorithms. The empirical contribution of proposed methodology is validated through 84 experimental combinations (7 balancing techniques × 3 algorithms × 4 datasets), which allows for detailed empirical comparisons of the effects of different resampling strategies and training algorithms. Empirical results demonstrated the feasibility and robustness of the methodology, as well as the ability to generalize across different domains. The proposed methodology can be reused and extended by design to include different data balancing techniques, algorithms and hyperparameter optimization strategies. The empirical evaluation metrics and particularly the overall score offer practical recommendations for companies developing machine learning models for churn prediction as part of their customer retention strategies. This justifies the practical contribution of the paper.

While the proposed approach demonstrated satisfactory performance across various datasets, techniques and models, we defined possible areas for future research on the topic. First, new machine learning algorithms can be explored, particularly deep learning neural networks and comparison with the achieved results. The set of evaluated data balancing techniques can be further enlarged to include other resampling methods like SMOTETomek, ADASYN with Random Forest, Majority Weighted Minority Oversampling Technique (MWMOTE) and others. Another important area of future research is focused on cost-sensitive learning and using profit-based metrics to achieve not only highly accurate models but also to maximize the business outcome. Acknowledging the importance of explainable AI (xAI), future research will integrate methods for model interpretation like SHAP, LIME, etc., to allow for better understanding of factors contributing to customer churn.

Another possible focus of future research is to develop churn prediction models to enable a more personalized approach and integration with different customer segmentation models. The addition of time-related features related to customer behavior over time may also be a promising area for providing more detailed knowledge on customer churn and allow for early detection.

Conclusion

The analysis suggests that among the various evaluated balancing techniques, SMOTE and SMOTEENN, particularly when used with XGBoost, have the best performance for churn prediction, making them the top choices for handling imbalanced data. Following closely is the combination of Random Forest with SMOTEENN, which also shows excellent results, especially in reducing false positives. Overall, we recommend evaluating different balancing techniques and choosing the most appropriate one with regard to the business requirements, dataset characteristics and the objectives of implementing machine learning models for churn prediction. It is also recommended to consider the misclassification costs to evaluate the costs, associated with false predictions.

Future research will explore the integration of these techniques with other machine leaning algorithms like deep learning neural networks, heterogenous ensembles and others to assess their performance across more datasets to validate their scalability and adaptability. To further enhance the empirical and theoretical contribution, new models will be added to the framework - for models' explanation, cost-sensitive learning, personalized churn prediction and time-series analysis on customer behavior over time.

Acknowledgements

This research is financed by the project "DNP-KS24-06-MLOCL - Machine learning for optimizing customer lifecycle".

References

Agarwal, P., Data Science Approaches for Churn Prediction, *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024, pp. 1-7, doi: 10.1109/ICCCNT61001.2024.10723983.

Ahmed, J., Younis, I., Sarwar, U., Ghaffar, R., & Ahmed, T. (2024). Leveraging Machine Learning Models for Customer Churn Prediction in Telecommunications: Insights and Implications. *VAWKUM Transaction on Computer Sciences*, 12(2), 16–27. doi: 10.21015/vtcs.v12i2.1904

Ako, R. E., Malasowe, B. O., Okpor, M. D., Akazue, M. I., Yoro, R. E., Ojugo, A. A., Setiadi, D. R. I. M., Odiakaose, C. C., Abere, R. A., Emordi, F. U., Geteloma, V., & Ejeh, P. O. (2024). Effects of Data Resampling on Predicting Customer Churn via a Comparative Tree-based Random Forest and XGBoost. *Journal of Computing Theories and Applications,* 2(1), 86–101. https://doi.org/10.62411/jcta.10562

Alhakim, M. F., Petchhan, J., & Su, S. (2024). Leveraging TabNet for Enhanced Customer Churn Prediction in the Telecommunication Industry. *2024 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), Taichung, Taiwan,* 2024, pp 717–718. doi:10.1109/icce-taiwan62264.2024.10674508

Atay, M. T., & Turanlı, M. (2024). Analysis of customer churn prediction using logistic regression, -nearest neighbors, decision tree and random forest algorithms. *Advances and Applications in Statistics*, 92(2), pp 147–169. doi:/10.17654/0972361725008

Awodele, O. et al., (2020). Enhanced Churn Prediction in the Telecommunication Industry. *International Journal of Innovative Research in Computer Science & Technology (IJIRCST),* 8(2).

Basha, H. S., Mittal, M., Chaithra, M. H., Kala, K., Patil, H., & Maranan, R. (2024). Predicting Customer Churn Rates in Service Providing Organizations Using Riemann Residual Neural Network with Snow Geese Algorithm, *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, Coimbatore, India. pp 1463–1469, doi: 10.1109/icoici62503.2024.10696036

Bhattacharjee, S., Thukral, U., & Patil, N. (2023). Early Churn Prediction from Large Scale User-Product Interaction Time Series. *arXiv.Org*, *abs/2309.14390*. doi: 10.48550/arxiv.2309.14390

Burez, J. and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction, *Expert Systems with Applications*, 36(3), pp. 4626–4636.

Demirbaga, Ü., Aujla, G.S., Jindal, A., Kalyon, O. (2024). *Machine Learning for Big Data Analytics*. In: Big Data Analytics. Springer, Cham. doi: 10.1007/978-3-031-55639-5_9

Denayer, J. (2024). Customer Churn Analysis in Telecom Industry using Kaplan–Meier and Cox Proportional Hazards Model. *Indian Scientific Journal Of Research In Engineering And Management*, 08(06), pp 1–5. doi:/10.55041/ijsrem36121

De Rainville, F.M. et al. (2012). DEAP: A Python framework for Evolutionary Algorithms. *GECCO'12 - Proceedings of the 14th International Conference on Genetic and Evolutionary Computation Companion*, pp. 85–92.

Frazier, P. I. (2018). A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*, Available at: https://arxiv.org/abs/1807.02811v1 (Accessed: 4 January 2024).

Gour, S. et al. (2024). Optimizing Hyperparameters in Neural Networks Using Genetic Algorithms. *Recent Advances in Computational Intelligence and Cyber Security*, pp. 106–116. Available at: https://doi.org/10.1201/9781003518587-9.

Haddadi, S., Farshidvard, A., Silva, F., JReis, J. & Reis, M., (2024). Customer churn prediction in imbalanced datasets with resampling methods: A comparative study. *Expert Systems with Applications*, Volume 246.

Hambali, M. A., & Andrew, I. (2024). Bank Customer Churn Prediction Using SMOTE: A Comparative Analysis. Qeios. https://doi.org/10.32388/h82xtw

Harshini, A., Ramya, N., Teja, B. S., Sandeep, C. S. S., & Shareefunnisa, S. (2024). Improving customer churn prediction accuracy: a SMOTE-based approach. *2024 8th International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, India, 2024, pp 215–222. https://doi.org/10.1109/icisc62624.2024.00044

Imani, M., Ghaderpour, Z., Joudaki, M., & Beikmohammadi, A. (2024). The Impact of SMOTE and ADASYN on Random Forest and Advanced Gradient Boosting Techniques in Telecom Customer Churn Prediction. *2024 10th International Conference on Web Research (ICWR)*, Tehran, Iran, Islamic Republic of, 2024, pp 202–209. https://doi.org/10.1109/ icwr61162.2024.10533320

Kun, L., Alli, H. & Rahman, K. (2024). GA optimization-based BRB AI reasoning algorithm for determining the factors affecting customer churn for operators. *Social Sciences & Humanities Open,* Volume 10.

Lalwani, P., Mishra, M. & Chadha, J. (2021). Customer churn prediction system: a machine learning approach. *Computing,* 271–294.

Ljubičić, K., Merćep, A. & Kostanjčar, Z.. (2023). Churn prediction methods based on mutual customer interdependence. *Journal of Computational Science,* 67(101940).

Loukili, M., Messaoudi, F. & El Ghazi, M. (2022). Supervised Learning Algorithms for Predicting Customer Churn with Hyperparameter Optimization. *International Journal of Advances in Soft Computing and its Applications,* 14(3), 49-63.

Matthews-El, T. & K., B., 2024. 14 Customer Retention Strategies That Work. [Online] Available at: (https://www.forbes.com/advisor/business/customer-retention-strategies/ [Accessed 23 02 2025].

Poudel, S., Pokharel, S. & Timilsina, M.. (2024). Explaining customer churn prediction in telecom industry using tabular machine learning models. *Machine Learning with Applications,* 17(100567).

Prabadevi, B., Shalini, B. & Kavitha, B. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks,* Volume 4, 145-154.

Pratheebha, T., Santhana Megala, S., & Indumathi, V. (2024). An Improved Churn Prediction Model using Relevance Feature Identification and Personalized Modeling Technique with Optimized Retention Strategies. *2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2024, pp. 1489-1496, doi: 10.1109/ICECA63461.2024.10800772.

Ren, H. (2025). Machine Learning-Based Prediction of Customer Churn Risk in E-commerce. *Advances in Economics, Management and Political Sciences*, 153(1), 47–52. https://doi.org/10.54254/2754-1169/2024.19473

Sánchez-Gutiérrez, M. E., & González-Pérez, P. P. (2023). Addressing the class imbalance in tabular datasets from a generative adversarial network approach in supervised machine learning. https://doi.org/10.1177/17483026231215186

Sarkar, T. D., Rahman, Md. A., & Karthikeyan, S. (2024). Analysis of Customer Churn for Telecom Company with SMOTE-ENN and Hyperparameter Tuning Randomized-SearchCV Technique in Advanced Machine Learning Technology. *2024 International Conference on Advancements in Power, Communication and Intelligent Systems (APCI)*, KANNUR, India, pp 1–6. doi:10.1109/apci61480.2024.10617375

Sathishkumar, V., Anilkumar, Ch., & Moorthy, U. (2023). Big Data Analytics and Implementation Challenges of Machine Learning in Big data. *Applied and Computational Engineering*, 2(1), 533–538. doi: 10.54254/2755-2721/2/20220584

Shumaly, S., Neysaryan, P. & Guo, Y., (2020). *Handling Class Imbalance in Customer Churn Prediction in Telecom Sector Using Sampling Techniques, Bagging and Boosting Trees.* Mashhad, IEEE, 082-087

Sharma, R. (2024). Customer Churn Analysis in Telecom Industry using Logistics Regression in Machine Learning with Kaplan–Meier and Cox Proportional Hazards Model. *Indian Scientific Journal of Research In Engineering And Management*. doi: 10.55041/ijsrem30745

Sun, Y., Liu, H. & Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Heliyon,* 9(2).

Thamizhselvi, D., Punitha, A., Kumar, R. S., Godson, V. R., Godson, D. R and Bharath, S.,

Churn Analysis For Restaurant, *2024 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, Chennai, India, 2024, pp. 1-6, doi: 10.1109/ICPECTS62210.2024.10780408.

Wagh, S., Andhale, A., Wagh, K.,Pansare, J., Ambadekar, S.& Gawande, S. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization,* Volume 14.

Wang, Ch., Rao, C., Hu, F., Xiao, X.& Goh, M. (2024). Risk assessment of customer churn in telco using FCLCNN-LSTM model. *Expert Systems with Applications,* 248(123352).

Yokoyama, A.M., Ferro, M. and Schulze, B. (2023). Multi-objective hyperparameter optimization approach with genetic algorithms towards efficient and environmentally friendly machine learning. *AI Communications*, 37(3), pp. 429–442.