Preprints.org

Article

Not peer-reviewed version

Evaluating an Ensemble of Random Forest and XGBoost with Gaussian Noise Upsampling Technique for Customer Churn Prediction

Mehdi Imani *

Posted Date: 30 October 2024

doi: 10.20944/preprints202410.2329.v1

Keywords: Customer Churn Prediction; XGBoost; Gaussian Noise Upsampling; Random Forest; Machine Learning

 doi:10.20944/preprints202410.2329.v1

Article

Evaluating an Ensemble of Random Forest and XGBoost with Gaussian Noise Upsampling Technique for Customer Churn Prediction

Mehdi Imani *, Danish Hashmi and Shalini Verma

Department of Computer and System Sciences, Stockholm University, 10691 Stockholm, Sweden; daha0543@student.su.se, shve0590@student.su.se
* Correspondence: m.imani@gmail.com,

Abstract: Customer churn is a critical challenge for subscription-based businesses, especially in telecommunications, where retaining customers is essential to maintaining profitability. This study investigates the efficacy of two ML models, XGBoost and Random Forest, for predicting customer churn using a publicly available telecommunications dataset. The dataset, characterized by imbalanced classes, presents a crucial challenge addressed by incorporating the Gaussian Noise Upsampling (GNUS) sampling technique. The study evaluates and compares the two models using essential performance indicators, including precision, recall, accuracy, F1-score, and ROC-AUC, both with and without GNUS sampling. The results indicate that while XGBoost initially outperforms Random Forest across most metrics, both models show improved recall after the GNUS application, particularly in identifying churn cases. However, this improvement in recall comes with a trade-off in precision and overall accuracy. The findings highlight the relevance of using appropriate sampling techniques to tackle class imbalance in churn prediction and provide valuable insights for developing proactive customer retention strategies.

Keywords—Customer Churn Prediction; XGBoost; Gaussian Noise Upsampling; Random Forest; Machine Learning

I. Introduction

Customer churn, losing customers or subscribers for any reason, is a significant challenge in the telecommunications, finance, and e-commerce industries. In today's competition, businesses must prioritize customer retention due to the high cost of acquiring new customers, which is often significantly higher than retaining existing ones. Studies consistently show that acquiring a new customer can be up to five times more expensive than retaining an existing customer, underscoring the importance of developing robust churn prediction models [1]. To address this challenging problem, Machine Learning (ML) has emerged as a powerful mechanism for predicting customer churn, enabling companies to proactively address customer dissatisfaction and implement retention strategies to reduce customer churn. ML algorithms can analyze customer data to identify patterns and behaviors that indicate churn scenarios. This ability to process complex datasets has made ML particularly effective in building predictive models. Various ML methods have been applied to churn prediction, including decision trees, logistic regression, support vector machines (SVM), and artificial neural networks (ANN). Each of these techniques offers distinct advantages. For example, decision trees are easy to interpret and implement, while more sophisticated methods like SVM and ANN can handle more complex patterns in customer behavior. In many cases, boosting algorithms like AdaBoost have been used to improve the performance of these models, mainly when dealing with imbalanced datasets [1].

The telecommunications industry has been the initial adopter of applying ML for churn prediction. As competition increases and customer satisfaction becomes more critical, telecom companies have utilized various ML techniques to predict and implement retention strategies to

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 October 2024 doi:10.20944/preprints202410.2329.v1

reduce churn. Research demonstrates that including features such as social network analysis (SNA) and other advanced feature engineering techniques can enhance the accuracy of churn prediction models [2]. These techniques leverage data from customer interactions, such as calls, text messages, and internet usage, to detect patterns that may indicate a likelihood of churn.

In this project, two ML classification techniques (XGBoost and Random Forest) will be used to assess the efficacy of predicting customer churn using a publicly available dataset from the telecommunication industry. The objective is to evaluate the performance of these techniques based on critical metrics, including precision, recall, accuracy, ROC-AUC, and F1-score, to identify the most effective approach for predicting churn.

II. Purpose of the Study

This research addresses the challenge of customer attrition, or churn, within the telecommunications industry. Customer churn has emerged as a critical issue for service providers, as the cost of acquiring new customers often surpasses that of retaining existing ones. To address this challenge, businesses increasingly turn to Customer Relationship Management (CRM) solutions to handle customer interactions and enhance retention strategies. These systems monitor customer engagement and use machine learning techniques to analyze and predict customer behavior, making them essential for understanding and preventing churn. This study aims to create practical solutions that help businesses classify customers into two groups: those likely to leave and those likely to stay, tackling a typical binary classification challenge.

The research focuses on two ML classification models, XGBoost and Random Forest, to evaluate their effectiveness in predicting customer churn using a publicly available dataset from the telecommunications sector.

This study thoroughly explores how XGBoost and Random Forest perform when applied to imbalanced datasets, a common issue in churn prediction. By analyzing the impact of these ML techniques on predictive accuracy, the research provides valuable insights into the application of ML in tackling the churn prediction challenge. The results will give deeper insights into the advantages and limitations of these approaches within the framework of customer retention strategies.

III. Related Work

Predicting customer churn has become increasingly important, especially in industries where retaining customers is critical to maintaining profitability. This is particularly true for sectors like telecommunications, banking, and retail. Various machine learning techniques and data mining methods have been used to improve the accuracy of churn prediction, helping businesses identify which customers are most likely to leave. These models use historical customer data, behavioral patterns, and social network influences to make accurate predictions. The literature highlights various approaches to churn prediction, including ensemble learning, handling imbalanced data, data preparation, hybrid techniques, and integrating social network analysis, each offering unique advantages.

A. Data Preparation Techniques

Data preparation plays a crucial role in improving model performance. A study [3] showed that preprocessing techniques like feature engineering, feature selection, and variable transformation could enhance model accuracy by up to 14.5%. The study concluded that a well-optimized logistic regression model could perform comparably to advanced ML models when data is preprocessed effectively. In a cross-company churn prediction context, the study [4] evaluated data transformation methods such as log transformation, Z-score normalization, and Box-Cox transformation. They found that log and Box-Cox transformations significantly improved model performance, while Z-score did not perform as well.

B. Addressing Class Imbalance

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 October 2024 doi:10.20944/preprints202410.2329.v1

One of the critical challenges in churn prediction is the imbalance between churners and non-churners. Study [5] tackled this issue by employing SMOTE (Synthetic Minority Over-sampling Technique) to oversample the minority class in a highly imbalanced Internet Service Provider dataset. The study compared models like XGBoost, AdaBoost, and KNN, finding that XGBoost offered the best performance in terms of both precision and recall. Study [6] also addressed class imbalance using cost-sensitive learning and random sampling, finding that under-sampling combined with weighted random forests improved performance, especially for datasets with extreme class imbalance. These findings also align with the study in [7], which introduced an improved Balanced Random Forest (IBRF) model to address the imbalance in churn datasets, further validating the effectiveness of cost-sensitive approaches.

C. ML Techniques for Churn Prediction

A variety of ML algorithms have been explored for churn prediction. Study [8] focused on Decision Trees and Support Vector Machines (SVM), demonstrating their effectiveness in churn prediction within the telecommunications industry. Similarly, a study in [9] conducted a comparative study on various ML models, finding that SVM-POLY combined with AdaBoost provided the highest accuracy and F-measure. A study in [10] compared several ML techniques, including Artificial Neural Networks (ANN) and Decision Trees, concluding that ANNs had superior performance, mainly when boosted, achieving an accuracy of 97.92%.

D. Ensemble Learning Techniques

Ensemble techniques have been widely adopted to improve the accuracy of churn prediction models by combining multiple weak learners into a more robust predictive model. Study [11] introduced nested ensemble learners using a combination of Boosting, Stacking, and Bagging techniques. The study demonstrated that hybrid models outperform traditional ensembles in churn prediction, achieving high accuracies on UCI and proprietary datasets. Study [12] explored hybrid resampling techniques like SMOTE-ENN combined with ensemble models such as XGBoost and LightGBM to handle imbalanced data. The study found that this approach significantly improved churn prediction, especially for minority classes such as churners, highlighting the power of ensemble methods in addressing class imbalance challenges.

E. Hybrid Learning Approaches

Hybrid models have gained popularity in churn prediction due to their ability to integrate the strengths of different models. A study in [13] proposed a hybrid classification algorithm that combines logistic regression with decision trees (LLM), showing improved predictive performance in terms of AUC and Top Decile Lift (TDL). A study in [14] proposed a method that estimates classifier certainty, integrating KNN and gradient-boosted trees (GBT), which was particularly useful for making accurate predictions in areas with higher certainty. This approach is promising for improving the robustness of hybrid models.

F. Rule-Based and Social Network Analysis Approaches

Social network analysis (SNA) has emerged as an innovative approach to churn prediction, particularly in telecommunications. A study in [15] proposed a model incorporating interpersonal influence through social networks. This study showed that integrating these relational aspects improved prediction accuracy significantly compared to traditional models. A study in [16] expanded on this by incorporating call detail records (CDRs) to map social networks, exhibiting that integrating relational learning with non-relational models significantly improves churn prediction accuracy.

G. Applications in Various Sectors

Churn prediction has been applied in sectors beyond telecommunications. A study in [17] explored the application of decision trees for churn prediction in China's Personal Handyphone System Service (PHSS). This study identified three experimental strategies to enhance prediction

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 October 2024 doi:10.20944/preprints202410.2329.v1

performance: altering sub-periods for training data, adjusting misclassification costs, and changing sample methods. The results demonstrated that decision trees, especially with optimal parameter settings like a sub-period length of 10 days and a misclassification cost of 1:5, achieved effective churn prediction even with limited variables such as Frequency of Use (FOU), Sphere of Influence (SOI), and Minutes of Use (MOU). [18] studied churn prediction in non-subscription industries, where defining churn is more complex. Their study showed that random forests reliably predicted churn in these non-traditional contexts. Similarly, Studies in [19] applied Artificial Neural Networks (ANNs) and decision trees in various industries, emphasizing the need for accurate data preprocessing techniques like feature selection and data balancing. Their findings reinforced the importance of proper data preparation in achieving robust predictions across diverse datasets.

In our prior research [20, 21], we examined the effects of various upsampling techniques on the performance of XGBoost and Random Forest models. However, these studies did not include the GNUS upsampling method, as the primary focus was applying SMOTE and ADASYN techniques.

IV. Method

This research aims to address the problem of customer attrition, commonly known as churn, in the telecommunications sector. Churn has become a significant issue, prompting service providers to focus more on retaining existing customers due to the high costs associated with acquiring new ones.

The research paper focuses on applying two ML techniques (XGBoost and Random Forest) on a publicly accessible dataset for predicting customer churn in the telecommunications sector. The main objective of these machine learning models is to predict and classify customers of churn and non-churn, presenting a binary classification task. This classification is crucial in the telecommunications sector, where customer retention is critical to maintaining revenue and reducing marketing costs, given the high costs associated with acquiring new customers.

Standard metrics such as Precision, Recall, Accuracy, F1-score, and ROC-AUC were employed for evaluation. These metrics provide a comprehensive review of the model's accuracy and ability to classify customers into churn or non-churn categories accurately.

The research contributes to the field by examining how different ML techniques can affect predictive accuracy when applied to imbalanced data. This comprehensive framework aims to provide subscription-based companies with practical tools for predicting customer churn, vital in the current data-centric business environment.

A. Training and Validation Process

This study uses k-fold cross-validation to evaluate the performance of the classifiers. However, one of the common challenges with this method is its handling of imbalanced datasets. In many cases, specific folds may not have enough examples from the minority class, which can skew the evaluation of the model's performance. To mitigate this, stratified k-fold cross-validation ensures that each fold maintains the same proportion of the minority and majority class instances, preserving the overall data distribution.

A vital aspect of this research is using the GNUS sampling technique (Gini Index, Nonuniform, Under-sampling, and Smoothing) to address the significant class imbalance in customer churn datasets. Before training the models, GNUS is applied to balance the dataset more effectively. This involves under-sampling the majority class but in a selective way that prioritizes the most relevant examples based on the Gini Index, ensuring that critical information is preserved during the sampling process. This approach helps retain essential instances from the majority class that contribute most to the model's learning, improving the model's predictive power on the minority class.

GNUS also incorporates nonuniform sampling, which focuses on making informed sampling decisions rather than random selection, further improving the model's ability to handle imbalanced data. Additionally, a smoothing step is applied during the process, reducing the chance of overfitting by minimizing the variance caused by imbalanced data, thus improving generalization to unseen data.

 doi:10.20944/preprints202410.2329.v1

All sampling adjustments through GNUS are performed exclusively within the training data after the train-test split or during the cross-validation process to prevent data leakage. This ensures that no information from the validation set is exposed to the model during training, producing a more accurate reflection of how the model performs in real-world scenarios.

This study rigorously evaluates the XGBoost and Random Forest models' ability to predict customer churn by combining GNUS with stratified cross-validation. This approach emphasizes maintaining a careful balance between correctly identifying customers likely to churn and achieving overall model accuracy.

B. Evaluation Metrics

Traditional accuracy metrics tend to fall short when dealing with imbalanced datasets, as they primarily focus on correctly classifying the majority class. This study emphasizes precision, recall, and the F1-score derived from the confusion matrix to provide a more accurate evaluation. These metrics are precious for assessing the model's performance in the minority class. Precision measures the accuracy of optimistic predictions, while recall captures the model's ability to identify all relevant instances. The F1-score is a comprehensive indicator of the model's effectiveness in balancing these two aspects.

In addition to threshold metrics, ranking metrics are employed to evaluate the models' performance across different decision thresholds. The Receiver Operating Characteristic (ROC) Curve is commonly used, as it illustrates the trade-off between the True Positive Rate and False Positive Rate at various threshold levels. The Area Under the ROC Curve gives a singular value that reflects the model's ability to distinguish between classes. However, when working with highly imbalanced data, the Precision-Recall Curve proves more suitable, as it concentrates on the model's capability to identify instances from the minority class correctly.

In summary, the study emphasizes the importance of using appropriate metrics for evaluating ML models in the context of customer churn prediction. This ensures that the evaluation accurately reflects the challenges posed by imbalanced data and focuses on the model's performance on the minority class.

V. Results

This section details the results of the simulations that we performed to evaluate XGBoost and Random Forest classification techniques to predict customer churn. This section contains two main parts: Setup and results.

A. Setup

The main goal of this study is to assess two ML techniques (XGBoost and Random Forest) used for customer churn prediction. Python programming language and libraries like Pandas, Numpy, etc., were used for simulations. The dataset used in this simulation came from Kaggle [22]. This dataset had 20 features divided into 4250 training and 750 testing records. This dataset was preprocessed to handle categorical variables, feature selection, and outlier removal before being used for model training and evaluation using key metrics such as precision, recall, F1-score, and ROC-AUC.

B. Results

The following table presents a comparative analysis of the Random Forest (RF) and XGBoost models on the customer churn dataset, with and without applying the GNUS sampling technique. The performance of the models is evaluated using several key metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

Table 1. Evaluation metrics.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 October 2024 doi:10.20944/preprints202410.2329.v1

Model Performance Metrics - Dataset 85 to 15

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RF-initial	0.917647	**0.951220**	0.435754	0.597701	0.832461
XGB-initial	**0.929412**	0.900901	0.558659	**0.689655**	**0.876076**
RF-gnus	0.906667	0.772727	0.474860	0.588235	0.813203
XGB-gnus	0.916078	0.839623	0.497207	0.624561	0.831490

Accuracy:

XGBoost-Initial demonstrated the highest accuracy, achieving 92.94%, slightly outperforming the RF-Initial model, which attained an accuracy of 91.76%. These figures suggest that XGBoost is better at correctly predicting churn and non-churn instances than the Random Forest model.

After applying GNUS sampling, both models experienced a slight reduction in accuracy. XGBoost-GNUS achieved an accuracy of 91.61%, and RF-GNUS gained 90.67%. The decrease in accuracy is expected due to the focus on improving predictions for the minority class (churn), which can slightly compromise overall accuracy.

Precision:

Precision is the proportion of correctly predicted churn cases out of all expected ones. The RF-initial model exhibited the highest precision, scoring 95.12%, indicating that the Random Forest model was very conservative in predicting churn, leading to fewer false positives.

However, this precision came at the expense of recall, as the RF model identified fewer actual churn cases. XGBoost-initial also performed well in precision, scoring 90.09%.

After applying GNUS, precision decreased for both models, reflecting a shift in focus from precision to recall. RF-GNUS showed an accuracy of 77.27%, and XGBoost-GNUS had a precision of 83.96%. This decrease suggests that the models predicted churn more frequently after GNUS sampling, resulting in more false positives and capturing more actual churn cases (improved recall).

Recall:

Recall, which measures the ability to identify actual churn cases correctly, was significantly lower in the initial models, particularly for the RF-initial model, which had a recall of 43.58%. This indicates that the initial Random Forest model struggled to identify many churn cases despite its high precision.

XGBoost-initial performed better, with a recall of 55.87%, demonstrating a more balanced approach between precision and recall.

After applying GNUS sampling, both models showed improvements in recall. The RF-GNUS model increased its recall to 47.49%, and XGBoost-GNUS improved to 49.72%. This indicates that GNUS helped both models capture more of the actual churn instances, which is crucial for customer churn prediction.

F1-Score:

The F1-Score was highest for the XGBoost-initial model, with a score of 68.97%. This reflects its better overall balance in identifying churn cases while maintaining precision.

The RF-initial model had a significantly lower F1-score of 59.77%, as it prioritized precision over recall, leading to fewer correctly predicted churn cases.

After GNUS sampling, the F1-scores of both models decreased slightly, with RF-GNUS scoring 58.88% and XGBoost-GNUS scoring 62.46%. This slight drop in F1-score reflects the trade-off between increasing recall and maintaining precision.

ROC-AUC:

The ROC-AUC metric, which measures the model's ability to distinguish between churn and non-churn classes, shows that XGBoost-initial outperformed the other models with an AUC of 87.61%, indicating that it was the best at distinguishing between the two classes.

The RF-initial model had an AUC of 83.25%, which is respectable but shows a weaker ability to separate the classes than XGBoost.

 doi:10.20944/preprints202410.2329.v1

After applying GNUS, both models experienced a slight reduction in their ROC-AUC scores. XGBoost-GNUS had an AUC of 83.15%, and RF (GNUS) had an AUC of 81.32%. This reduction aligns with the focus on improving recall at the expense of some precision and overall classification ability.

In summary, the XGBoost model consistently outperformed the Random Forest model across most metrics before and after applying GNUS sampling. XGBoost-initial had the best balance between precision and recall and the highest F1-score and ROC-AUC. However, after applying GNUS sampling, both models improved in the recall, making them more effective at identifying churn cases, though with some precision and overall accuracy trade-offs. These results highlight the challenge of balancing precision and recall in customer churn prediction and demonstrate the benefits of using techniques like GNUS sampling to improve the identification of minority class instances.

Figures 1, 2, 3, and 4 display the ROC curve diagram for Random Forest and XGBoost with and without the GNUS sampling technique.

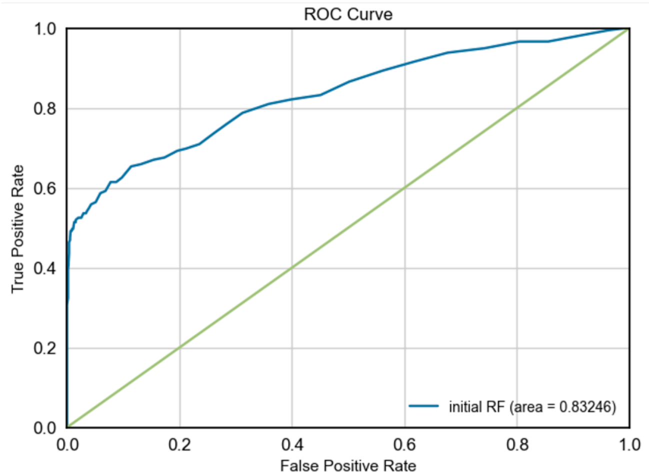

Figure 1. ROC curve diagram for RF-initial.

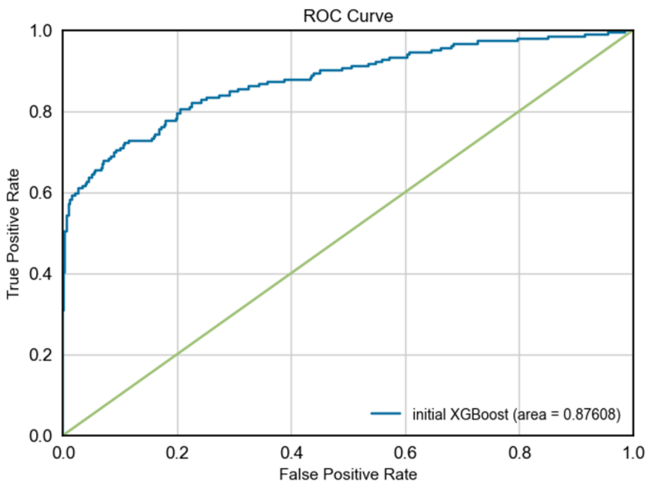

Figure 2. ROC curve diagram for XGBoost-initial.

 doi:10.20944/preprints202410.2329.v1

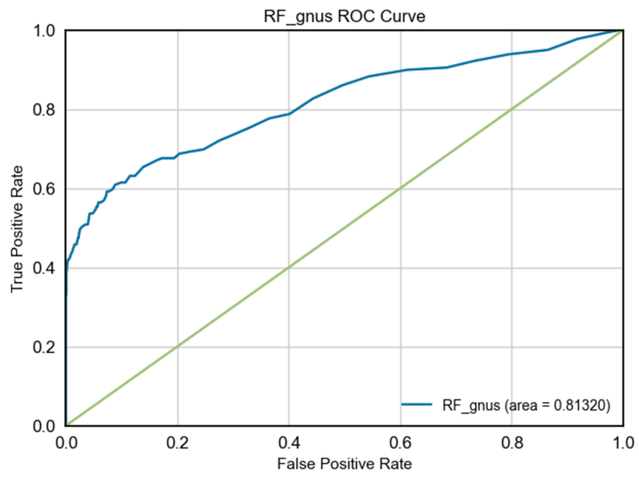

Figure 3. ROC curve diagram for RF-GNUS.

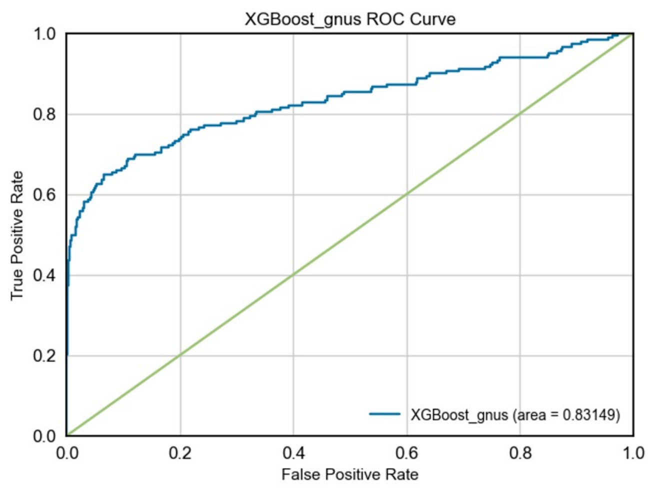

Figure 4. ROC curve diagram for XGBoost-GNUS.

In summary, the XGBoost model consistently outperformed the Random Forest model across most metrics before and after applying GNUS sampling. XGBoost-initial had the best balance between precision and recall and the highest F1-score and ROC-AUC. However, after applying GNUS sampling, both models improved in recall, making them more effective at identifying churn cases, though with some precision and overall accuracy trade-offs. These results highlight the challenge of balancing precision and recall in customer churn prediction and demonstrate the benefits of using techniques like GNUS sampling to improve the identification of minority class instances.

Conclusions

Customer churn poses a substantial problem for businesses, especially in highly competitive industries like telecommunications. The cost of acquiring new customers far outweighs retaining existing ones, making accurate churn prediction essential for minimizing revenue loss and improving customer retention strategies.

This study aimed to compare the performance of two prominent machine learning models, XGBoost and Random Forest, for predicting customer churn using an imbalanced dataset. The results indicate that XGBoost outperformed Random Forest in several key metrics, including accuracy, F1-score, and ROC-AUC, with XGBoost-initial achieving an accuracy of 92.94% and an F1-score of 68.97%. After applying the GNUS sampling technique, the Random Forest model saw a notable improvement in recall, increasing from 43.58% to 47.49%, indicating better identification of churn cases. In contrast, XGBoost's recall decreased slightly from 55.87% to 49.72%, showing a trade-off between recall and precision after applying GNUS. Precision for both models decreased post-GNUS, with Random Forest-GNUS and XGBoost-GNUS prioritizing recall at the cost of misclassifying some non-churn customers.

Overall, GNUS sampling effectively improved Random Forest's ability to capture churn cases, making it more suitable for scenarios where identifying churners is critical. However, for XGBoost, GNUS led to a slight decline in performance on key metrics. The study highlights the challenge of balancing precision and recall, emphasizing that different models and techniques should be selected based on specific business goals.

For future work, further exploration into other advanced sampling techniques or hybrid approaches could be valuable in mitigating the trade-offs between precision and recall. Additionally, incorporating features such as social network analysis or more granular customer behavior data could enhance the model's predictive capabilities. This research contributes to developing robust machine-learning solutions for customer churn prediction and offers practical insights for industries looking to optimize customer retention strategies.

References

1. Vafeiadis, Thanasis & Diamantaras, Kostas & Sarigiannidis, G. & Chatzisavvas, Konstantinos. (2015). A Comparison of Machine Learning Techniques for Customer Churn Prediction. Simulation Modelling Practice and Theory. 55. 10.1016/j.simpat.2015.03.003.
2. Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in the big data platform. J Big Data 6, 28 (2019).
3. Kristof Coussement, Stefan Lessmann, Geert Verstraeten, A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry, Decision Support Systems, Volume 95, 2017, Pages 27-36, ISSN 0167-9236.
4. Adnan Amin, Babar Shah, Asad Masood Khattak, Fernando Joaquim Lopes Moreira, Gohar Ali, Alvaro Rocha, Sajid Anwar, Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods, International Journal of Information Management, Volume 46, 2019, Pages 304-319, ISSN 0268-4012.
5. D. Do, P. Huynh, P. Vo, and T. Vu, "Customer churn prediction in an internet service provider," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 2017, pp. 3928-3933.
6. J. Burez, D. Van den Poel, Handling class imbalance in customer churn prediction, Expert Systems with Applications, Volume 36, Issue 3, Part 1, 2009, Pages 4626-4636, ISSN 0957-4174.
7. Yaya Xie, Xiu Li, E.W.T. Ngai, Weiyun Ying, Customer churn prediction using improved balanced random forests, Expert Systems with Applications, Volume 36, Issue 3, Part 1, 2009, Pages 5445-5449, ISSN 0957-4174.
8. Jadhav, Rahul & Pawar, Usharani. (2011). Churn Prediction in Telecommunication Using Data Mining Technology. International Journal of Advanced Computer Sciences and Applications. 2. 10.14569/IJACSA.2011.020204.
9. T. Vafeiadis, K.I. Diamantaras, G. Sarigiannidis, K.Ch. Chatzisavvas, A comparison of machine learning techniques for customer churn prediction, Simulation Modelling Practice and Theory, Volume 55, 2015, Pages 1-9, ISSN 1569-190X.
10. Wouter Verbeke, David Martens, Christophe Mues, Bart Baesens, Building comprehensible customer churn prediction models with advanced rule induction techniques, Expert Systems with Applications, Volume 38, Issue 3, 2011, Pages 2354-2364, ISSN 0957-4174.
11. Ahmed, Mahreen & Afzal, Hammad & Siddiqi, Imran & Amjad, Muhammad & Khurshid, Khawar. (2020). Exploring nested ensemble learners using overproduction and choosing an approach for churn prediction in the telecom industry. Neural Computing and Applications. 32. 10.1007/s00521-018-3678-8.
12. Kimura, Takuma. (2022). Customer Churn Prediction with Hybrid Resampling and Ensemble Learning.. 1-23.
13. Arno De Caigny, Kristof Coussement, Koen W. De Bock, A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, European Journal of Operational Research, Volume 269, Issue 2, 2018, Pages 760-772, ISSN 0377-2217, https://doi.org/10.1016/j.ejor.2018.02.009.
14. Adnan Amin, Feras Al-Obeidat, Babar Shah, Awais Adnan, Jonathan Loo, Sajid Anwar, Customer churn prediction in telecommunication industry using data certainty, Journal of Business Research, Volume 94, 2019, Pages 290-301, ISSN 0148-2963, https://doi.org/10.1016/j.jbusres.2018.03.003.
15. Xiaohang Zhang, Ji Zhu, Shuhua Xu, Yan Wan, Predicting customer churn through interpersonal influence, Knowledge-Based Systems, Volume 28, 2012, Pages 97-104, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2011.12.005.

 doi:10.20944/preprints202410.2329.v1

16. Wouter Verbeke, David Martens, Bart Baesens, Social network analysis for customer churn prediction, Applied Soft Computing, Volume 14, Part C, 2014, Pages 431-446, ISSN 1568-4946.
17. L. Bin, S. Peiji, and L. Juan, "Customer Churn Prediction Based on the Decision Tree in Personal Handyphone System Service," 2007 International Conference on Service Systems and Service Management, Chengdu, China, 2007, pp. 1-5.
18. Jennifer Karlberg, Maja Axén. (2020). Binary Classification for Predicting Customer Churn. Department of Mathematics and Mathematical Statistics at Umeå University.
19. Shaaban, Essam & Helmy, Yehia & Khedr, Ayman & Nasr, Mona. (2012). A Proposed Churn Prediction Model. International Journal of Engineering Research and Applications (IJERA. 2. 693-697.
20. Imani, Mehdi, et al. "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). IEEE, 2024.
21. Imani, Mehdi, and Hamid Reza Arabnia. "Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: a comparative analysis." Technologies 11.6 (2023): 167.
22. Rinichristy, "Customer Churn Prediction 2020," Kaggle, Dec. 12, 2022.https://www.kaggle.com/code/rinichristy/customer-churn-prediction-2020