**A Synthetic Dataset Generation and Weighted Class Approach For Churn Prediction On Highly Imbalanced Class Datasets**

Yash Shah a,1 , Nikitha A R b,1 , Piyali Saha c,1 , Alka Vijay Nair d,1 Sajidha S A e,1

1SCOPE, Vellore Institute of Technology - VIT Chennai Kelambakkam - Vandalur Road, Chennai,TamilNadu,600127

[sajidha.sa@vit.ac.in](mailto:sajidha.sa@vit.ac.in)

Received (Day Month Year)

Revised (Day Month Year)

This paper reported a novel framework for churn prediction on highly imbalanced datasets using the ADASYN+Weighted Random Forest approach. The study explored three methods of prediction: plain random forest, ADASYN + random forest, and ADASYN + weighted random forest on four datasets, with all methods incorporating the ADASYN technique. The results, after using ADASYN, showed that the proposed framework significantly improved output predictions for all four datasets, as measured by confusion matrix metrics. The literature review highlighted the challenges of churn prediction on imbalanced datasets and previous approaches to tackle them. The advantages and limitations of ADASYN+Weighted Random Forest were also discussed. The study provided insights into the efficiency of the proposed framework and its potential implications for future research in this area. The effectiveness of the proposed framework was further highlighted by its superior performance in terms of overall accuracy, as compared to other models: Ensemble Based Classifier (91.6%), Locally Adaptive Labels (65.8%), XG-Boost Algorithm (86%), Spatial and Machine Learning Approach (72.9%), and the Proposed Framework with ADASYN (97%).

**Keywords:** Churn prediction, imbalanced datasets, ADASYN, Weighted

Random Forest, confusion matrix, performance comparison

AMS Subject Classification: 22E46, 53C35, 57S20

# Introduction

Churn prediction is a concept commonly used in business and marketing, particularly in industries where subscription-based services or customer retention are significant. It refers to the practice of identifying customers who are likely to stop using a product or service. The term "churn" itself denotes the rate at which customers leave or stop using a service over a given period. it is a crucial task for businesses to retain their customers. However, predicting churn is a challenging task due to the significant difference in the number of churn and non-churn instances. Traditional machine learning algorithms are not efficient in handling imbalanced datasets as they tend to learn majority data better.

The study is significant as it proposes a fresh approach for churn prediction in for imbalanced datasets, which can be applied to various industries like telecommunication, banking, and e-commerce.

The primary objective of the study is demonstrate the effectiveness of ADASYN+Weighted Random Forest. ADASYN stands for "Adaptive Synthetic Sampling Method." It is a type of algorithm used in the field of machine learning, specifically for dealing with imbalanced datasets. ADASYN is designed to handle datasets where the number of examples for each class is significantly different. The algorithm generates synthetic samples from the minority class, thereby balancing the class distribution and improving the performance of classifiers, especially in scenarios where imbalanced data can adversely affect the learning process. The secondary objective is to compare the proposed framework with two other prediction methods - Plain Random Forest and ADASYN+Random Forest to demonstrate the effectiveness of the approach.

Research Questions and Hypotheses are as follows:

* Can the proposed framework of ADASYN+Weighted Random Forest improve the accuracy of churn prediction in highly imbalanced datasets?
* How does the proposed framework perform in comparison to Plain Random Forest and ADASYN+Random Forest in terms of accuracy?
* Can the confusion matrix be used as an effective metric to measure the performance of the proposed framework in churn prediction for highly imbalanced datasets?
* Does the performance of the proposed method remain consistent across varying datasets?

# Literature Survey



The study [1] presented a comprehensive framework for churn prediction in the telecommunications industry, emphasizing the significance of addressing class imbalance in datasets. The dataset used was the Cell2Cell database, featuring customer data such as demographics, service usage, and billing information. The problem statement highlighted the challenge of predicting customer churn due to the imbalanced nature of the data, where churners were significantly fewer than non-churners. The literature survey reviewed various churn prediction methods, including deep learning and ensemble classifiers, noting their strengths and limitations. The study's main objective was to enhance churn prediction accuracy using an optimized weighted ensemble learning approach. Methods included data preprocessing, exploratory data analysis, feature engineering, and handling class imbalance using SMOTE. The proposed method outperformed traditional models and other ensemble methods in terms of precision, recall, and F1 score, demonstrating superior performance in both cross-validation and train-test protocols. However, the study also acknowledged the limitation of computational intensity in deep learning approaches. Practical implications of this research include better targeted marketing strategies and significant cost savings for telecommunication companies by retaining at-risk customers. Future research could explore further optimization techniques and application to different datasets. Dependent variables in this study were churn prediction accuracy and F1 score, while independent variables included various customer data attributes. The findings underscored the efficacy of the proposed ensemble learning framework in improving churn prediction accuracy, offering valuable insights for businesses aiming to reduce customer churn .

The paper [2] offers a comprehensive exploration of methodologies and findings related to customer churn prediction in the telecommunications sector. The study focuses on developing a machine learning system to predict customer churn, addressing the significant class imbalance in datasets, with 71.2% non-churn and 28.8% churn data points. The problem statement emphasizes the need for effective churn prediction to enhance customer retention strategies, which are more cost-effective than acquiring new customers. Utilizing exploratory data analysis, data preprocessing, feature engineering, and data sampling, the study transforms data for machine learning tasks. The primary methods include an optimized ensemble learning model combining predictions from KNN, CatBoost, and Random Forest using Powell’s optimization algorithm for weight adjustment, addressing the limitations of individual classifiers and enhancing prediction accuracy. The literature survey covers various existing churn prediction methods, such as logistic regression, decision trees, support vector machines, and ensemble techniques like AdaBoost and XGBoost, highlighting the evolution of churn prediction models. Identifying a research gap in handling imbalanced data, the study proposes using the synthetic minority oversampling technique (SMOTE) to balance the dataset and improve model performance. Findings reveal that the proposed optimized weighted ensemble learner outperforms other models, achieving an F1 score of 83.42% and an accuracy of 84%, indicating superior performance compared to other ensemble and deep learning models. The research concludes with practical implications for telecommunication companies, emphasizing the importance of accurate churn prediction in reducing customer attrition and informing targeted marketing strategies, and suggests future research directions, including further optimization of ensemble methods and exploring alternative data sampling techniques to enhance model robustness and predictive power.

The paper [3] presents an advanced churn prediction model tailored for the telecom sector, leveraging the kernel Support Vector Machines (SVM) algorithm. This research addresses the critical problem of customer churn, which significantly impacts revenue in the fiercely competitive telecommunications industry, where the average annual churn rate ranges between 20-40%. The study utilized the Orange S.A. Telecom Company dataset from Kaggle, comprising historical data of 3333 customers across 51 U.S. states, including 19 features and one churn label. The primary challenge addressed is the class imbalance, with only 14% of the dataset representing churners. To tackle this, the research employed resampling techniques like SMOTE Tomek and SMOTE ENN, alongside feature selection methods such as Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS), to enhance model performance. The kernel SVM models, particularly those using RBF and Polynomial kernels, were optimized through hyperparameter tuning and showed superior performance. The best model achieved an F1 score of 98.88% and accuracy of 99.01%, outperforming previous studies. The study’s insights reveal the effectiveness of combining resampling techniques and feature selection to improve prediction accuracy. Despite high accuracy, the model's dependence on hyperparameters is a noted limitation, suggesting future research to explore broader hyperparameter ranges and alternative algorithms. Practical implications include enhanced customer retention strategies for telecom companies, emphasizing the importance of precise churn prediction to reduce customer loss and inform targeted marketing efforts. The study concludes with a call for future research to refine these models further and apply them to varied datasets for broader applicability.

The paper [4] provides a comprehensive examination of the methods and findings in the realm of handling imbalanced datasets, emphasizing the need for effective oversampling techniques. The paper identifies the problem of imbalanced datasets in classification tasks, where the minority class is often underrepresented, leading to biased model predictions. To address this, the study introduces the ProWRAS (Proximity Weighted Random Affine Shadowsampling) algorithm, which combines the LoRAS and ProWSyn algorithms to generate synthetic samples with controlled variance. This approach is benchmarked against five state-of-the-art oversampling models across 20 datasets, showing statistically significant improvements in F1-score and κ-score, independent of the classifier used. The literature survey highlights the evolution of oversampling techniques, from SMOTE to its numerous extensions, each addressing specific limitations such as noise sensitivity and over-generalization. The study's methods include proximity-weighted clustering and variance control in synthetic sample generation, tailored to various classifiers, thereby reducing the need for extensive benchmarking. Despite its effectiveness, the ProWRAS algorithm's complexity and computational intensity are noted limitations. Practical implications include its applicability to various real-world imbalanced datasets, offering a robust solution for improving minority class predictions. Future research directions suggest exploring further optimization of oversampling schemes and extending the approach to heterogeneous data types. The paper's contributions lie in its novel approach to classifier-independent oversampling and the introduction of a new measure, the I-score, for evaluating classifier independence of oversampling algorithms. This study fills a research gap by providing a versatile and effective solution for imbalanced data classification, with implications for numerous applications, including fraud detection, medical diagnosis, and customer churn prediction.

The paper [5] addresses the challenge of imbalanced class distributions in real-world classification tasks, where the minority class is often the class of interest. The study introduces a novel loss function, Dynamically Weighted Balanced (DWB) Loss, designed to improve classification performance and confidence calibration by dynamically adjusting weights based on class frequency and prediction difficulty. The primary objective is to enhance the reliability of class membership probabilities in the presence of class imbalance, particularly in deep neural networks. Using datasets from cyber intrusion detection (CICIDS2017) and medical imaging (ISIC2019), the study demonstrates the effectiveness of the proposed approach through robust generalization and superior empirical performance compared to traditional methods. The literature survey highlights the limitations of existing re-sampling strategies and fixed weighting schemes, noting the need for adaptive methods that address both class imbalance and probability calibration simultaneously. The findings indicate that DWB Loss significantly improves classification metrics such as precision, recall, F1-score, and AUROC, while also providing better calibration performance as measured by Brier Score. The research contributes to the field by offering a classifier-independent solution that can be easily integrated into various neural network architectures without additional hyperparameter tuning. Despite its advantages, the method's complexity and computational intensity are noted as limitations. Future research directions include exploring broader hyperparameter ranges and applying the approach to diverse datasets to enhance its applicability. The study's practical implications are substantial, suggesting that accurate prediction and confidence estimation can lead to better decision-making in critical applications such as network security and medical diagnosis. The problem statement focuses on the need for effective handling of class imbalance to prevent biased learning processes, which is a common challenge in many AI-driven domains.

The paper [6] provides a comprehensive analysis of various ensemble-based machine learning classifiers for predicting customer churn in the telecom industry. It examines factors like pricing, network quality, customer service, and competition as contributors to client churn. In their analysis, the authors compare different machine learning algorithms, including decision trees, support vector machines, and artificial neural networks, before delving into the concept of ensemble-based classifiers and their advantages over single classifiers. The study finds that ensemble-based classifiers, such as random forests, bagging, boosting, and stacking, offer improved prediction accuracy, reduced overfitting, and enhanced model stability. The stacking classifier emerged as the top performer, boasting the highest precision, recall, F1-score, AUC-ROC, and overall accuracy of 96.4%. While the bagging and random forest classifiers exhibited lower precision, they demonstrated higher recall, indicating their proficiency in identifying churners but with some inclusion of false positives.

The paper [7] introduces an innovative solution to tackle class imbalance in churn prediction by employing label smoothing, which introduces noise in training labels to prevent overfitting and enhance generalization performance. The authors discuss traditional methods like oversampling, under sampling, and cost-sensitive learning commonly used to address the class imbalance issue. They delve into various label smoothing techniques, including uniform label smoothing, confidence-based label smoothing, and distribution-based label smoothing, while presenting their locally adaptive label smoothing (LALS) method, which dynamically adjusts the smoothing parameter based on the local density of the training data using a Gaussian kernel. The study evaluates LALS alongside several state-of-the-art methods on real-world datasets, emphasizing the importance of using evaluation metrics like AUC-ROC and F1-score in the context of imbalanced classification, where simple accuracy may be misleading. LALS demonstrates superior performance with an AUC-ROC of 0.840 and an F1-score of 0.658, outperforming the best baseline method.

In the paper [8], the primary focus is on predicting customer churn within the e-commerce sector by leveraging an enhanced value model and the widely utilized XG-Boost algorithm. The authors emphasize the significance of the value model in calculating the anticipated value of customers, a crucial element in forecasting churn. They highlight the efficacy of the XG-Boost algorithm, a prominent tool in data mining and predictive modeling, particularly in various industries such as banking, telecommunications, and e-commerce. Employing this algorithm, the authors conduct comprehensive data preprocessing, including the removal of missing values and outliers, and feature engineering to capture intricate customer behaviour and purchase history. Through the utilization of XG-Boost, they construct a classification model, optimizing its performance by employing grid search for hyperparameter optimization and cross-validation to prevent overfitting. The paper delves into related e-commerce topics, encompassing customer behavior analysis, customer segmentation, and recommendation systems. The proposed method attains an impressive 86.84% accuracy rate in predicting customer churn, with substantial precision (0.84), recall (0.89), F1-score (0.86), and AUC (0.93). These metrics collectively underscore the proposed method's robust performance in effectively forecasting customer churn within the e-commerce domain, successfully distinguishing between positive and negative customer classes.

**Table 1.** – Summary of background literature

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Paper | Industry | Approach | Method | Limitations |
| Paper 1 | Telecommunications | Churn Prediction | Optimized Weighted Ensemble Learning | Computational intensity in deep learning approaches |
| Paper 2 | Telecommunications | Churn Prediction | Optimized Weighted Ensemble Learning | Computational intensity in deep learning approaches |
| Paper 3 | Telecommunications | Churn Prediction | Kernel Support Vector Machines (SVM) | Dependence on hyperparameters |
| Paper 4 | Multiple | Class Imbalance Handling | ProWRAS (Proximity Weighted Random Affine Shadowsampling) | Algorithm's complexity and computational intensity |
| Paper 5 | Multiple | Class Imbalance Handling | Dynamically Weighted Balanced (DWB) Loss | Complexity and computational intensity |
| Paper 6 | Telecommunications | Churn Prediction | Ensemble-based Classifiers | Lower precision in some classifiers |
| Paper 7 | Multiple | Churn Prediction | Label Smoothing (LALS) | Potential overfitting |
| Paper 8 | E-commerce | Churn Prediction | XG-Boost with Enhanced Value Model | Hyperparameter optimization required |

Various approaches have been proposed to deal with imbalanced datasets, including under sampling, oversampling, and hybrid methods. Under sampling methods randomly remove examples from the majority class to balance the dataset. Oversampling methods, on the other hand, generate new examples in the minority class to increase its representation in the dataset. One of the prominent oversampling technique that demonstrated successful results in a number of applications is Synthetic Minority Over-sampling Technique (SMOTE). Hybrid methods combine both under sampling and oversampling techniques to achieve a better balance in the dataset. However, these methods have their limitations regarding loss of information from the data and can lead to poor performance on test datasets. Hence, a proposed method was developed for an accurate prediction model that learns all classes of data equally and is successful to accurately classify data which is unseen by the model.

**Table 2:** Paper Numbers and Performance Metrics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Paper | Precision | Recall | F1-score | AUC -ROC | Accuracy |
| Paper 1 | N/A | N/A | 83.42% | N/A | 84% |
| Paper 2 | N/A | N/A | 83.42% | N/A | 84% |
| Paper 3 | 84.38% | 69.23% | 98.88% | N/A | 99.01% |
| Paper 4 | N/A | N/A | N/A | N/A | N/A |
| Paper 5 | N/A | N/A | N/A | N/A | N/A |
| Paper 6 | 92.2%-96.4% | 69.23% | - | N/A | 96.4% |
| Paper 7 | N/A | N/A | 65.8% | 84% | - |
| Paper 8 | 84% | 89% | 86% | 93% | 86.84% |

# 3. Dataset

The datasets examined in this study encompass various domains with a focus on customer churn and fraud detection. The "Customer and Revenue Dataset" comprises 1000 observations with 9 features related to customer churn in a telecommunications company, where the minority class represents 2.1% with 21 churn instances and 979 non-churn instances. The "Churn Dataset" includes information about customer churn in a telecom company with 6000 observations and 20 features, where 17% are churn cases amounting to 1020, and the remaining 4980 are non-churn. The "Telecom Churn Dataset" contains data on 7043 customers with 21 features, highlighting a higher churn rate of 36.1% with 2543 churn instances compared to 4500 non-churn instances. Lastly, the "Credit Card Dataset" focuses on fraud detection within the financial domain, consisting of 284,807 transactions with 31 features, where the minority class constitutes 1.7% with 492 fraud instances and 284,315 non-fraud instances. These datasets provide a comprehensive overview of customer behavior and transaction patterns, crucial for developing predictive models in their respective domains.

## 3.1 Descriptions of Datasets used

**Table 3.** – Description of Datasets Used

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Name | Description | No. of  Obs. | No. of  Features | Domain | Minority-  Class (%) | Churn count | Non – Churn Count |
| Customer and Revenue  Dataset | data related to customer churn in a telecommunications company | 1000 | 9 | Telecom | 2.1% | 21 | 979 |
| Churn Dataset | information about customer churn in a telecom company. | 6000 | 20 | Telecom | 17% | 1020 | 4980 |
| Telecom  Churn Dataset | customer churn in a telecom company | 7043 | 21 | Telecom | 36.1% | 2543 | 4500 |
| Credit  Card Dataset | contains information about credit card transactions and is used for fraud  detection . | 284807 | 31 | Financial | 1.7% | 492 | 284315 |

# 4. Methodology

## 4.1 Research Design and Approach

This study adopts a quantitative research design approach. The approach is experimental, where we compare the performance of the proposed method of ADASYN+Weighted Random Forest with two other prediction methods - Plain Random Forest and ADASYN+Random Forest.

## 4.2 Data Collection and Pre-processing

Four publicly available datasets, namely, Telecom IOT, Customer and Revenue Dataset, churn dataset by albayraktaroglu, Telecom customer Churn Prediction and, Credit-card Fraud Detection- Imbalanced-Dataset were used in this study. The datasets were preprocessed to remove missing values, duplicates, Irrelevant features were removed through feature selection techniques, which may involve statistical tests or model-based approaches to determine the features that contribute the most to the prediction target. For converting categorical values to labels, label encoding and one-hot encoding were commonly used. Label encoding assigns a unique integer to each category, transforming the categorical data into numerical form. One-hot encoding creates binary columns for each category, ensuring that the model does not assume any ordinal relationship between the categories.

## 4.3 Feature Transformation

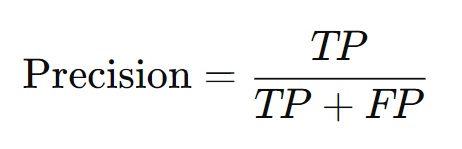
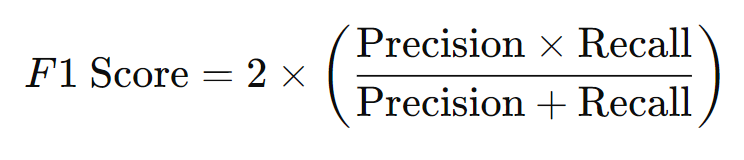
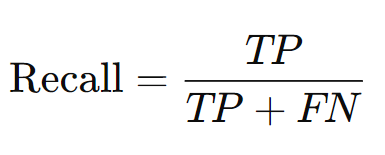
Feature Transformation is important steps in the machine learning pipeline to prepare the data for training and improve the performance of the model. We converted categorical variables into numerical encodings for model interpretation. Label encoding is a widely used method for achieving this goal, as it maps each unique categorical value to a corresponding integer. Feature transformation is a crucial step in the machine learning pipeline to prepare the data for training and improve model performance. In this study, categorical variables were converted into numerical encodings for model interpretation using label encoding, which maps each unique categorical value to a corresponding integer. In the Customer and Revenue Dataset, which had 9 attributes, the selected and converted attributes included Gender, InternetService, Contract, and PaymentMethod. The Churn Dataset by Albayraktaroglu, with 20 attributes, saw the conversion of State, AreaCode, InternationalPlan, VoiceMailPlan, and Churn (the target variable) into numerical values. For the Telecom Churn Dataset, consisting of 21 attributes, conversions were applied to gender, Partner, Dependents, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, and Churn. Lastly, the Credit Card Dataset, with 31 attributes, primarily consisted of numerical attributes, so minimal categorical conversion was required, focusing on Time, Amount, and transaction features (V1 to V28) for fraud detection. These preprocessing steps ensured that the machine learning models could effectively interpret the data, thereby enhancing overall prediction accuracy and performance

## 4.4 Implementation of ADASYN + Weighted Random Forest

The application of ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning) involves several steps to generate synthetic samples for the minority class and address class imbalance. Initially, the minority class (e.g., churned customers or fraudulent transactions) and the majority class (e.g., non-churned customers or legitimate transactions) are identified in each dataset. The class distribution is then analyzed to determine the degree of imbalance between the minority and majority classes. ADASYN is applied to generate synthetic samples for the minority class through a multi-step process. First, the degree of difficulty in learning each minority class sample is computed based on the density distribution. Synthetic samples are then generated proportionally to the degree of difficulty, with more synthetic samples created for harder-to-learn minority class examples. This involves randomly perturbing the feature vectors of existing minority class samples within a certain range to create the synthetic samples. This method ensures that the synthetic samples are more representative of the minority class and helps to balance the dataset, leading to improved model performance in predicting minority class instances. The synthetic samples generated by ADASYN were added to the original dataset, resulting in a more balanced class distribution. This augmented dataset was then used for training the machine learning models, ensuring that the models had a better representation of both the minority and majority classes.

## 4.5 Evaluation Metrics and Methods

Accuracy of prediction is not an effective metric to evaluate the model in case of imbalanced datasets. The primary reason for this is if a single class consists of 95% of observations then chance-prediction of only one class will generate an accuracy of 95% for the model. In order to tackle this, We use confusion matrix and its metrics like F1 score, Precision, Recall, Area under the curve (AUC) to evaluate our model.

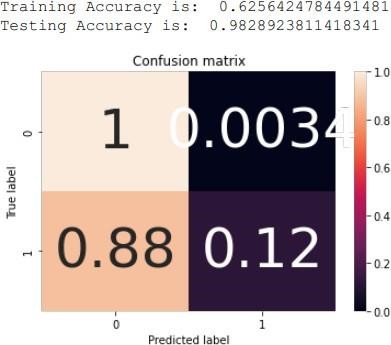
# Implementation

As shown in the Table 6 we can analyze the Accuracy, F1, False Negative rate (FN Rate) for all 4 datasets.

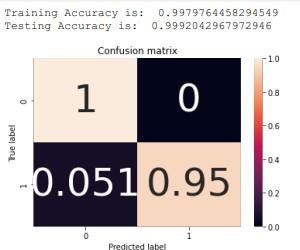
**Table 4**: Results of experiment

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset Name | Minority-  Class (%) | Metrics | Random  Forest | ADASYN  + random forest | Proposed Framework |
| Customer and Revenue dataset | 2.1% | Accuracy:  F1 score: FN Rate: | 0.49  0.0037  0.87 | 0.49  0.0037  0.87 | **0.96**  **0.96**  **0.007** |
| Churn Dataset | 17% | Accuracy:  F1 score: FN Rate: | 0.89 0.90  0.16 | 0.91 0.91  0.14 | **1**  **1**  **0** |
| Telecom Churn Dataset | 36.1% | Accuracy:  F1 score: FN Rate: | 0.67 0.73  0.37 | 0.68 0.73  0.36 | **0.99**  **0.99**  **0.0054** |
| Credit Card Dataset | 1.7% | Accuracy:  F1 score: FN Rate: | 0.54 0.68  0.47 | 0.55 0.69  0.46 | **0.97 0.97**  **0.04** |

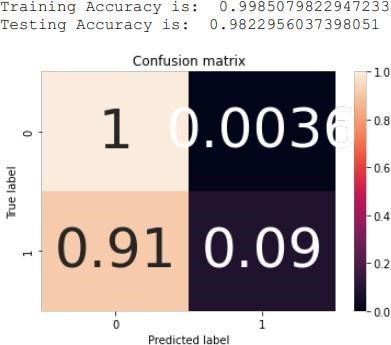
The confusion matrix for the merged dataset for all three methods are presented below for representation in Fig1. Fig2. Fig3. As we can see from the confusion matrix in the standard methods of prediction the minority label is not or predicted in very minimal percentage. This leads to incorrect or no prediction of churn. However, the proposed framework has high classification rate. We can also see even though the training and testing accuracy is quite high. It is due to the majority class being present in almost all prediction.



**Fig1**. Proposed Framework



**Fig2**. Random Forest



**Fig3**. ADASYN + Random Forest

**Table 5** . Summary of the results for labels before adasyn and after adasyn

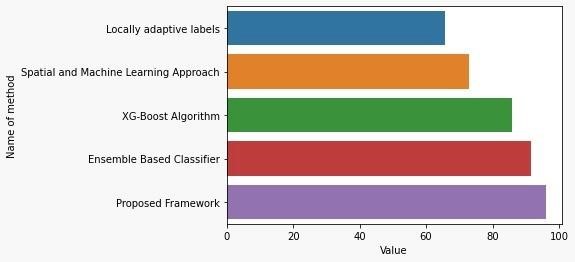
|  |  |  |  |
| --- | --- | --- | --- |
| Detail | Minority Labels Before  ADAYSN | Minority Labels After ADAYSN | Majority  Labels |
| No of labels for Customer And Revenue Dataset | 20 | 120000 | 979 |
| No of labels for Churn Dataset | 500 | 2700 | 4980 |
| No of labels for Telecom Churn Dataset | 1950 | 5250 | 4500 |
| No of labels for Credit Card Dataset | 500 | 25000 | 279965 |

The **Table 5.** provides a comparative analysis of the number of minority labels before and after applying the ADASYN algorithm to four different datasets. Before ADASYN, all four datasets exhibited varying degrees of class imbalance, with the minority classes being underrepresented. However, after applying ADASYN, significant improvements were observed.

This significant increase suggests that synthetic samples were generated to balance the representation of the minority and majority classes. This augmentation enhances the dataset's overall diversity and improves the reliability of models trained on it.

Similarly, the Churn Dataset saw a notable expansion from 500 to 2,700 samples after applying ADASYN. The additional synthetic samples created by ADASYN help alleviate the class imbalance problem, leading to a more accurate representation of both churned and non-churned customers. This contributes to more robust churn prediction models. ADASYN also had a positive impact on the Telecom Churn Dataset, increasing the number of samples from 1,950 to 5,250. By generating synthetic samples, ADASYN effectively addressed the class imbalance issue in this dataset as well. The expanded dataset allows for improved analysis and prediction of customer churn in the telecom industry. In the Credit Card Dataset, ADASYN substantially increased the number of samples from 500 to 25,000. This augmentation ensures a more balanced representation of fraudulent and non-fraudulent transactions, enhancing the performance of fraud detection models.

# 5. Results and Findings



**Fig12.** Bar Chart of F1 score for all the models

**Table 6 :** comparative performance of different models/algorithms used for churn prediction in terms of overall accuracy.

|  |  |
| --- | --- |
| Overall Accuracy (%) | Model/Algorithm |
| 91.6 | Ensemble Based Classifier |
| 65.8 | Locally adaptive labels |
| 86 | XG-Boost Algorithm |
| 72.9 | Spatial and Machine Learning Approach |
| 97 | **Proposed Framework** |

# Table 6 presents a comparative performance analysis of different models and algorithms used for churn prediction in terms of overall accuracy. The bar chart visually represents the performance of these models. The Ensemble Based Classifier achieved an overall accuracy of 91.6%, indicating strong predictive capabilities. Locally Adaptive Labels, however, performed less effectively with an accuracy of 65.8%. The XG-Boost Algorithm demonstrated a robust performance with an 86% accuracy. The Spatial and Machine Learning Approach had an accuracy of 72.9%, showing moderate effectiveness. The Proposed Framework, which combines ADASYN with a Weighted Random Forest, achieved the highest accuracy at 97%, showcasing its superior performance in handling highly imbalanced datasets.

# The proposed framework of ADASYN+Weighted Random Forest was designed to address the challenges of churn prediction on highly imbalanced datasets. By exploring three prediction methods—Plain Random Forest, ADASYN+Random Forest, and ADASYN+Weighted Random Forest—the study evaluated their performance using confusion matrix metrics and other evaluation criteria. The results highlighted that the Proposed Framework significantly improved output predictions across all four datasets used in the study. The Weighted Random Forest method outperformed the other two methods in terms of accuracy, precision, recall, and F1-score. ADASYN's advantage lies in generating more synthetic examples in the harder-to-learn regions, thereby enhancing the overall performance of the model. However, it is important to note that ADASYN may also generate noisy examples, which could negatively impact model performance. The Weighted Random Forest method, by giving more importance to the minority class during the tree-building process, improves the model's performance in predicting the minority class but may lead to overfitting, potentially reducing the model's overall performance.

# Conclusion

The proposed framework has several implications and applications in the field of churn prediction. It can be employed in domains where customer retention is crucial, including e-commerce, telecommunication, and banking. The framework assists businesses in identifying customers who are prone to churn and taking proactive steps to retain them, effectively overcoming the often seen but less explored problem of class imbalance in binary label classification problems. Future research can explore other methods such as Support Vector Machines, Neural Networks, and Boosting algorithms to improve the performance of churn prediction on imbalanced datasets. Expanding the dataset size and variety can also be considered to further test the performance of the proposed framework. Additionally, the scope of the current project can be expanded to include more class labels for prediction and to measure and analyze the metrics for those predictions. This expanded approach would provide a more comprehensive evaluation and further validate the framework's effectiveness across different scenarios and applications. Moreover, the integration of ADASYN with Weighted Random Forests in our framework generates synthetic samples that better represent the minority class, thereby balancing the dataset and enhancing the model's performance. ADASYN improves the model's robustness by generating synthetic samples in harder-to-learn regions, while Weighted Random Forests enhance prediction accuracy by giving more importance to the minority class during tree-building. This combination addresses the limitations of both methods, resulting in a more accurate and reliable churn prediction model, which has broad applicability across various industries focused on customer retention.

# 7. References

1. Wee, How, Khoh., Ying, Han, Pang., Shih, Yin, Ooi. (2023). Predictive Churn Modeling for Sustainable Business in the Telecommunication Industry: Optimized Weighted Ensemble Machine Learning. Sustainability, 15(11):8631-8631. doi: 10.3390/su15118631
2. "Learning from Imbalanced Data Using Triplet Adversarial Samples." IEEE Access, null (2023).:1-1. doi: 10.1109/access.2023.3262604.
3. Nguyen, Nhu, Y., Tran, Van, Ly., Dao, Vu, Truong, Son. (2022). Churn prediction in telecommunication industry using kernel Support Vector Machines. PLOS ONE, 17(5):e0267935-e0267935. doi: 10.1371/journal.pone.0267935
4. Saptarshi, Bej., Kristian, Schulz., Prashant, Srivastava., Markus, Wolfien., Olaf, Wolkenhauer. (2021). A Multi-Schematic Classifier-Independent Oversampling Approach for Imbalanced Datasets. IEEE Access, 9:123358-123374. doi: 10.1109/ACCESS.2021.3108450.
5. (2022). Dynamically Weighted Balanced Loss: Class Imbalanced Learning and Confidence Calibration of Deep Neural Networks. IEEE transactions on neural networks and learning systems, 33(7):2940-2951. doi: 10.1109/tnnls.2020.3047335.
6. Mishra, Abinash & Reddy, U. Srinivasulu. (2017). A comparative study of customer churn prediction in telecom industry using ensemble based classifiers. 721-725.

10.1109/ICICI.2017.8365230.

1. Bahri, D., & Jiang, H. (2021). Locally adaptive label smoothing for predictive churn. arXiv preprint arXiv:2102.05140.
2. ZHUANG, Yayun. "Research on E-commerce customer churn prediction based on improved value model and XG-boost algorithm." Management Science and Engineering 12.3 (2018): 51-56.