Improved Decision Tree, Random Forest, and XGBoost Algorithms for Predicting Client Churn in the Telecommunications Industry

Mohamed Ezzeldin Saleh, Nadia Abd-Alsabour
Department of Computer Science, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

***Abstract*—Traditional machine learning models, especially decision trees, face great challenges when applied to high-dimensional and imbalanced telecommunication datasets. The research presented in this paper aims to enhance the performance of traditional Decision Tree (DT), Decision Tree with grid search (DT+), random forest (RF), and XGBoost (XGB) models. This is accomplished by augmenting them with robust preprocessing techniques, as well as optimizing them through grid search. We then evaluated how well the enhanced models can accurately predict customer churn and compared their performance metrics in detail. We utilized a dataset derived from the benchmark Cell2Cell dataset by applying combined preprocessing methods including KNN imputation, normalization, and resampling with SMOTE Tomek to address class imbalance. The findings reveal that XGBoost outperformed all other models with an accuracy of 0.82, demonstrating strong precision, recall, and F1 scores. RF also delivered robust results, achieving an accuracy of 0.82, benefiting from its ensemble nature to improve generalization and reduce overfitting.**

***Keywords*—Churn prediction; decision trees; grid search; random forest; XGBoost**

I. Introduction

The rapid evolution of the telecommunications industry has been marked by significant technological advancements and massive competition, leading to a saturated market where customer retention has become a critical challenge. As new telecom providers emerge, often offering specialized services at competitive prices, established firms must adopt more sophisticated strategies to maintain their market share. Customer churn has emerged as a significant concern in this highly competitive environment. Retaining existing customers is more cost-effective than acquiring new ones, but it is also crucial for maintaining steady revenue streams and long-term business growth [1]-[2].

Machine Learning (ML) has transformed the telecom industry by providing advanced tools for analyzing large datasets and predicting customer behavior. ML algorithms, particularly those focused on customer churn prediction (CCP), enable telecom companies to implement proactive retention strategies by accurately identifying at-risk customers [3]-[4]. However, the high dimensionality and imbalance inherent in telecom datasets pose significant challenges to traditional ML models like Decision Trees (DT). While these models are popular due to their simplicity and interpretability, they often suffer from overfitting and biased predictions when applied to complex, high-dimensional data [5]. The need for robust preprocessing techniques, such as imputation, normalization, and resampling, is essential to address these limitations and improve the predictive performance of ML models [6]-[7].

Despite advances in ML, gaps remain in optimizing churn prediction models for real-world applications. Current methodologies often struggle with imbalanced datasets, leading to skewed predictions that fail to capture minority class churners effectively. Ensemble models such as Random Forest (RF) and XGBoost (XGB) offer improved generalization and accuracy but require careful tuning of hyperparameters to maximize their effectiveness. Moreover, the lack of standardized preprocessing pipelines and scalable solutions limits the broader adoption of these methods in the telecom industry.

This study addresses these challenges by proposing a systematic approach to enhance the performance of DT, RF, and XGB models. The research emphasizes integrating advanced preprocessing techniques such as KNN imputation, normalization, and SMOTE Tomek resampling—with hyperparameter optimization using grid search. This approach seeks to mitigate the impact of imbalanced datasets and improve the robustness of predictive models for CCP.

The key objectives of this work are to:

- Develop a DT+ model optimized through grid search to address the limitations of traditional DT models.
- Compare the outcomes of DT+, RF, and XGB models in predicting customer churn, focusing on precision, accuracy, F1-score, and recall.
- Investigate the impact of preprocessing techniques, including imputation, normalization, and resampling, on the performance of these models.

By addressing these objectives, this research aims to contribute to the development of scalable, reliable, and interpretable models for customer churn prediction, offering actionable insights for the telecommunications industry to retain customers and reduce churn rates effectively.

The literature review is depicted in the following section. Section III introduces the proposed work with enhanced algorithms and its pseudocode. Hyperparameter optimization is presented in Section IV, the performance metrics are

addressed in Section V. Results and discussion are in Section VI. Section VII gives detail about the CCP performance. The closing section addresses the real-world significance in Section VIII, with limitations, conclusion, and the future scope of research of this research in Section IX.

II. Literature Review

CCP has become an important area of concern in the telecommunications industry, prompting extensive research on the effectiveness of various ML models. Among the most studied algorithms are DT, RF, and XGB, each offers unique strengths in enhancing predictive accuracy and model stability.

The study by [8] introduced a smart hybrid scheme that combined clustering and classification algorithms. Their results demonstrated that a stacking-based ensemble model combining k-medoids, Gradient Boosted Tree (GBT), DT, RF, and Deep Learning (DL) achieved the highest accuracy of 96%, highlighting the potential of hybrid models.

The study in [9] explored the use of advanced machine learning methods, particularly focusing on RF optimized by Grid Search and a low-ratio undersampling strategy. Their findings showed that the RF-GS-LR model achieved near-perfect accuracy on the applied datasets, underscoring the importance of hyperparameter optimization and sampling techniques in improving churn prediction models.

The study in [10] provided a comprehensive analysis of integrated algorithms, including enhanced Random Forest and XGBoost models.

Studies by [11] and [12] emphasized the role of big data platforms, ensemble methods, and attribute selection in enhancing the accuracy and stability of churn prediction models. Using techniques such as SMOTE and Edited Nearest Neighbor (ENN) for data balancing and ensemble procedures like bagging and boosting has significantly improved model performance.

DT methodologies are highly regarded for their straightforwardness and interpretability. The research in [13] demonstrated that the DT models can achieve 3% higher accuracy than more complex models, such as random forests, under certain conditions. However, DT models are susceptible to overfitting and need to be enhanced through hyperparameter tuning methods such as grid search. The study in [9] illustrated how grid search optimization could significantly improve DT's accuracy and stability, especially when combined with controlled undersampling strategies.

RF is an ensemble method built on multiple DTs and is recognized for its robustness and ability to handle large datasets. The study in [14] highlighted RF's high accuracy of 95% when feature engineering techniques were applied, underscoring the importance of preprocessing steps [14]. In spite of its preferences, RF's complexity can be a restriction. However, studies have shown that when RF is regularized or combined with techniques like up-sampling and Edited Nearest Neighbor (ENN), it can achieve exceptionally high accuracies, reaching up to 99.09% [11], [15].

XGB, a gradient-boosting technique, has received widespread popularity for its outstanding usefulness in classification activities, particularly in churn prediction. Studies have revealed that XGB outperformed other ensemble algorithms, including Adaboost and CatBoost, especially when coupled with grid search cross-validation for hyperparameter tuning. XGB's capacity to handle sparse data and mitigate overfitting makes it highly effective for complex datasets [16]-[17]. The study in [18] further validated XGB's efficacy, achieving a 97% accuracy rate on the Cell2Cell dataset.

Integrating decision trees, random forests, and XGBoost algorithms alongside advanced optimization techniques like Grid search provides a comprehensive and robust approach to churn prediction. Studies by [19] emphasized the advancements in predictive power achieved through the combination of feature engineering, ensemble methods, and hyperparameter optimization, which significantly improved the accuracy, stability, and generalization across diverse telecom datasets.

Despite the advancements in CCP models, there are still several gaps that need to be addressed. Traditional decision tree models, while effective, are prone to overfitting and require optimization techniques like grid search to achieve optimal performance. Existing literature has shown improvements through these techniques, but further exploration is needed to address limitations in model interpretability and scalability [9], [13]. Random forest models, though robust, present challenges in terms of complexity and computational cost [11], [14]. The reliance on feature engineering to achieve high accuracy suggests the need for more efficient methods to handle raw data effectively. Additionally, although XGBoost shows superior performance, its susceptibility to overfitting and the need for extensive hyperparameter tuning suggest opportunities for further research on more generalized models. Furthermore, while recognized as crucial, hyperparameter optimization and sampling techniques require deeper investigation to develop standardized methodologies that can be applied to different datasets and industries [16]-[17].

III. Proposed Work

This section depicts the study details, pre-processing, and methodologies employed.

The study used the comprehensive dataset Cell2Cell, which contains client behavior attributes such as personal information, utilization patterns, client interactions, demographic details, billing data, and value-added services. These properties provide a solid foundation for developing and validating machine learning models [8], [20].

The research process, illustrated in Fig. 1, is structured into distinct phases, beginning with data preprocessing, which is crucial to ensuring the accuracy and reliability of the models. KNN Imputation (Mean/Median) addresses lost values within the dataset. This is followed by normalization utilizing MMADN Min-Max Scaling, and class imbalance is managed with SMOTE Tomek [20].

The preprocessed dataset is then split into training and testing parts in an 80-20 ratio, facilitating a robust evaluation of the models' performance. The DT and DT+ models serve as the baseline, with DT+ incorporating enhancements such as optimized hyperparameters identified through grid search techniques. RF, known for its ensemble approach, is also applied to improve the prediction accuracy and generalization further. Finally, XGBoost, renowned for its efficiency and scalability, is utilized, leveraging grid search for hyperparameter tuning.

The models were implemented on Google Colab, leveraging its computing resources for efficient processing. The analysis underscores the importance of systematic preprocessing and model enhancement in achieving high accuracy in churn prediction. By comparing the performance metrics of the DT, DT+, RF, and XGB models, this study provides empirical evidence for the effectiveness of advanced classification algorithms in CCP.

The preprocessing phase is essential to ensuring accurate predictions for customer churn models. Missing data is addressed using KNN imputation and median/Mean imputation techniques, preserving the dataset integrity and reducing bias [20]-[21]. Normalization follows, employing a combination of MMADN and Min-Max Scaling, which standardizes features while managing outliers effectively [22], [23], [24], [25]. To address the class imbalance, the SMOTE Tomek was utilized [18], [20], [25]. These preprocessing steps ensure that the data is optimally prepared for applying the DT, DT+, RF, and XGB models, leading to improved prediction accuracy and model robustness.

The proposed approach involves analyzing customer churn using the following four different algorithms:

- Traditional Decision Tree (DT)
- Decision Tree with Grid Search (DT+)
- Random Forest (RF)
- XGBoost (XGB)

A. Traditional Decision Tree (DT)

The Traditional Decision Tree model is valued for its effortlessness and intuitive interpretability. It works by recursively partitioning the dataset based on feature values, creating a tree-like structure where nodes represent decision rules and leaves denote class labels. This model excels at handling non-linear relationships, making it a popular choice for classification tasks. Nevertheless, DTs are inclined to overfit, particularly with complex or noisy data, necessitating careful tuning of parameters like max_depth, min_samples_split, and min_samples_leaf to enhance generalization.

B. Decision Tree with Grid Search (DT+)

The Decision Tree with grid search model refines the traditional Decision Tree by incorporating advanced techniques to mitigate overfitting and improve predictive accuracy. DT+ is designed to balance model complexity and interpretability, making it particularly suitable for applications where both robust performance and transparency in decision-making are crucial. By optimizing hyperparameters, DT+ effectively captures meaningful patterns from data, addressing the limitations of the traditional DT model.

DT+ leverages a few essential parameters to boost performance:

- Max_depth: Restricts the tree's depth, averting the model from getting too complicated and prone to overfitting
- Min_samples_split: Defines the minimum number of samples required to split a node, reducing the risk of creating insignificant splits.
- Criterion: The choice of Gini impurity or entropy when splitting a node directly affects the quality of the formed decision boundary.
- Pruning Techniques: These are applied post-training to remove branches that do not provide significant power in classifying the target variable, further reducing overfitting.

The implementation of DT+ begins with training an initial Decision Tree model using default parameters to establish a baseline. This model is then evaluated using accuracy metrics, a classification report, and a confusion matrix. To enhance the traditional DT model, grid search performs hyperparameter tuning to identify the best parameter settings from a predefined distribution. The tuned model is then evaluated on the test set, with pruning techniques applied to ensure that the model is not only accurate but also generalizable. This systematic approach ensures that the DT+ model outperforms the standard DT by avoiding overfitting and improving decision-making transparency.

Despite its enhancements, DT+ faces several challenges:

- Overfitting: Although it can be mitigated via pruning and parameter tuning, the risk of overfitting still exists, especially when the tree becomes too complex.
- Sensitivity to Data Variability: Like the traditional DT, DT+ can be sensitive to small changes in the dataset, which might lead to significant variations in the tree structure.
- Computational Complexity: Including advanced techniques such as hyperparameter tuning and pruning increases the computational burden, particularly with large datasets and extensive parameter grids.

The Pseudocode for the DT+ model development is:

Input: dataset.csv, Output: Model evaluation metrics, plots, comparison CSV.

- Import Libraries
- Load Dataset
- Data Preparation
- Initial Model Training with Default Parameters
- Hyperparameter Tuning using Grid search

- Comparison of Parameters and Accuracy
- Visualization Based on Performance Metrics

C. Random Forest (RF)

The Random Forest algorithm addresses single Decision Trees' limitations, particularly their susceptibility to overfitting. It is an ensemble learning strategy that boosts classification performance by consolidating the predictions of a number of Decision Trees, each trained on distinct portions of the data and attributes. This ensemble approach improves generalization, stability, and accuracy, making RF a powerful tool for complex classification tasks such as CCP. RF's ability to handle large datasets with high dimensionality and provide feature importance estimation further strengthens its applicability in various domains.

RF relies on several key parameters to optimize its performance:

- n_estimators: Determine the no. of decision trees in the ensemble. Increasing this number generally improves accuracy but increases computational costs.
- max_attribute: Determine the maximum no. of attributes considered for partitioning at every node, which provides randomness & diversity amongst the trees, boosting the model's robustness.
- Tree-Specific Parameters: These include parameters like max_depth, min_samples_split, and criterion, similar to those in Decision Trees but applied collectively across all trees in the ensemble.

The development of RF starts with creating an ensemble of decision trees. Every tree is trained on a bootstrap example of the data, with a random portion of attributes chosen for every partition. The last prediction is obtained by aggregating the predictions from all trees, regularly by means of majority voting. Hyperparameters such as the no. of trees (n_estimators), maximum features (max_features), and tree-specific parameters are fine-tuned to optimize the model's performance. Grid search allows for efficient hyperparameter tuning, ensuring that the model generalizes well to unseen data.

RF offers several advantages over single-decision trees. One of its main merits is enhanced generalization. By averaging the predictions of numerous trees, RF diminishes the risk of overfitting, which typically enhances execution on novel data. In addition, the strength of RF comes from the diversity amongst its constituent trees, making it less sensitive to noise & variability in the data. Another advantage is its ability to provide estimates of feature importance, which can be valuable for interpreting the model's decisions.

However, RF also has its drawbacks. The ensemble nature of RF demands significantly more computational resources than single Decision Trees, thus increasing the computational complexity. Moreover, while RF's accuracy is superior, its predictions are less interpretable due to the complexity involved in aggregating the outputs of multiple trees. Reduced comprehensibility hinders comprehension of how the model arrived at its conclusions.

D. XGBoost (XGB)

XGBoost (Extreme Gradient Boosting) is employed in CCP to leverage its superior performance in handling complex data interactions and minimizing errors through iterative refinement. Differentiated from conventional ensemble strategies, XGBoost develops trees sequentially, with each novel tree rectifying the errors caused by the past ones [10], [16]. This iterative strategy, integrated with gradient descent optimization, permits XGBoost to accomplish large accuracy & strength in classification issues. Its ability to incorporate regularization techniques makes it particularly effective in preventing overfitting [17], [19].

XGBoost's performance is susceptible to its hyperparameters, and tuning these parameters is crucial to prevent overfitting and ensure robust performance. Grid search efficiently explores a wide range of hyperparameter combinations, allowing for a thorough yet computationally feasible optimization process [16]-[17].

XGBoost's performance is highly dependent on several key parameters:

- Learning Ratio: Controls the commitment of each tree to the last model. A lower learning rate requires more boosting rounds but can lead to a better generalization.
- Greatest Depth (max_depth): Constrains the depth of each tree, adjusting model complexity with overfitting risk.
- No. of Boosting Rounds (n_estimators): Determines the no. of trees to be included successively. More trees can capture more patterns, but this may increase the risk of overfitting.

Implementing XGBoost begins with importing the necessary libraries. The data is split into training and testing parts, similar to the process used for RF. A parameter grid is set, and Grid search is utilized to seek the ideal hyperparameters efficiently. This approach ensures that the model achieves the best possible performance while avoiding overfitting. The best model identified through Grid search is then trained on the full training set and evaluated on the test set to validate its accuracy and generalization capabilities.

While XGBoost offers several advantages, including superior accuracy and the ability to handle missing data, it also presents challenges:

- Computational Complexity: XGBoost's iterative approach and need for extensive hyperparameter tuning can increase computational costs.
- Overfitting: Despite its regularization techniques, XGBoost can still overfit, particularly on small or noisy datasets, if not properly tuned.
- Interpretability: The complexity of the model can make it challenging to interpret, especially when compared to simpler models like decision trees.

The detailed Pseudocode for the XGB model development has been described below.

Input: dataset.csv, Output: Model evaluation metrics, plots

- Import Libraries
- Load Dataset
- Data Preparation
- Split Data
- Hyperparameter Tuning with Grid search
- Train the Best Model
- Evaluate the Model
- Visualization Based on Performance Metrics

IV. Hyperparameter Optimization

Hyperparameter tuning was conducted using grid search to optimize the performance of the predictive models. This process systematically evaluated a range of hyperparameter combinations to identify the best configuration for each model. The optimal hyperparameters for the DT+, RF, and XGBoost models are summarized below.

A. Improved Decision Tree (DT+)

The DT+ model, an optimized version of the traditional Decision Tree, was tuned to improve its performance by leveraging grid search. The best-performing hyperparameter combination for DT+ included:

- class_weight: {0: 1, 1: 5}
- criterion: entropy
- max_depth: 70
- max_features: None
- min_samples_leaf: 1
- min_samples_split: 2
- splitter: random

This configuration balanced the dataset effectively, reducing overfitting and improving decision-making across deep trees.

B. Random Forest (RF)

The Random Forest model, an ensemble technique, was optimized to maximize generalization and reduce variance. The best hyperparameter combination for RF was:

- bootstrap: False
- class_weight: {0: 1, 1: 2}
- criterion: gini
- max_depth: None
- max_features: log2
- min_samples_leaf: 1
- min_samples_split: 2
- n_estimators: 120

This configuration emphasized utilizing a larger number of estimators while balancing the dataset using class_weight, enhancing the model's robustness.

C. XGBoost (XGB)

The XGBoost model demonstrated its strength in handling high-dimensional and imbalanced datasets with the following optimized hyperparameters:

- learning_rate: 0.05
- max_depth: 50
- n_estimators: 100
- subsample: 0.8

This combination provided a balance between the learning rate and the depth of the trees, enabling the model to refine predictions iteratively while avoiding overfitting.

D. Impact of Hyperparameter Optimization

The tuning process significantly contributed to the improved performance of the models, as demonstrated in the results:

- DT+: Showed notable consistency in metrics, particularly recall, due to balanced class weights and randomized splitting criteria.
- RF: Achieved strong generalization with an accuracy of 0.82 and a ROC-AUC of 0.87, leveraging its optimal tree-based ensemble design.
- XGB: Delivered the best overall performance with an ROC-AUC of 0.88, benefiting from gradient boosting and iterative error correction.

These hyperparameter combinations underscore the importance of systematic optimization in achieving reliable and accurate predictions for CCP.

V. Performance Metrics

To evaluate the predictive accuracy of our models in churn prediction, we utilize key metrics from the confusion matrix, which categorizes predictions into four essential types: true positives, true negatives, wrong positives, and wrong negatives. These metrics are critical for assessing the effectiveness of our classification algorithms [22], [23], [24], [25].

- F-measure: Balances precision and recall by calculating their harmonic mean, offering a single metric to evaluate overall model performance.
- Precision: Valuates the accuracy of figuring out the extent of true positives with regard to all discovered churners.
- Recall: Measures the model's ability to correctly identify actual churners by calculating the proportion of true positives out of all actual churners.
- Accuracy: Reflects the overall correctness of the model by measuring the ratio of correct predictions to total predictions.

VI. Results and Discussion

This section compares the DT, DT+, RF, and XGB models to assess their effectiveness in predicting customer churn in the telecommunications industry. The performance of each model is evaluated based on accuracy, precision, recall, F1-score, and ROC-AUC to provide a clear understanding of their relative strengths.

The effectiveness of the classification models was assessed using the prepared Cell2Cell dataset. Preprocessing steps included imputation, normalization, and resampling techniques, ensuring the dataset was adequately prepared for model evaluation. As shown in Table I, ensemble methods (RF and XGB) significantly outperformed single-tree models (DT and DT+).

TABLE I. The Performance of Utilized Models

Models	Accuracy	Precision	Recall	F1-Score	ROC
DT	0.77	0.78	0.77	0.77	0.74
DT+	0.77	0.77	0.77	0.77	0.73
RF	0.82	0.83	0.82	0.81	0.87
XGB	0.82	0.82	0.82	0.81	0.88

- DT Model: The DT model reported moderate performance with an accuracy of 0.77, precision of 0.78, recall and F1-score of 0.77, and ROC-AUC of 0.74. These metrics highlight the baseline capability of a traditional decision tree structure.
- DT+ Model: With grid search hyperparameter optimization, the DT+ model achieved similar results, with accuracy, precision, and recall, an F1-score of 0.77 each, and a ROC-AUC of 0.73. The enhancements ensured consistent performance but did not significantly outperform the traditional DT.
- RF Model: The RF model demonstrated substantial improvement across all metrics. It achieved an accuracy of 0.82, precision of 0.83, recall of 0.82, and F1-score of 0.81. Its ensemble approach effectively leveraged imputed data and the SMOTE Tomek resampling method, resulting in a high ROC-AUC value of 0.87.
- XGBoost Model: XGBoost delivered the highest overall performance, achieving an accuracy of 0.82, precision, recall, and F1-score of 0.82 each, and a ROC-AUC of 0.88. The advanced optimization of gradient boosting and integration of regularization features contributed significantly to this outcome.

Fig. 2 to 7 provide classification reports and confusion matrices for the DT+, RF, and XGBoost models, illustrating their predictive performance.

The results highlight the advantages of ensemble methods like RF and XGBoost over single-tree models in churn prediction. The XGBoost model's superior performance can be attributed to its gradient-boosting framework, which iteratively refines predictions, effectively capturing complex patterns in the data. This aligns with findings from prior studies that emphasize the efficacy of gradient-boosting techniques for high-dimensional datasets. The RF model's robust results further underscore the value of ensemble techniques in enhancing generalization and reducing overfitting. The use of SMOTE Tomek in preprocessing was critical in addressing the class imbalance, as evident in the improved recall and precision scores for both RF and XGBoost. However, the limited impact of this technique on DT and DT+ indicates that more sophisticated models are better equipped to exploit balanced datasets.

Fig. 1. CCP performance of ML models.

```
Accuracy of the model:
0.7699539712075213
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.83      0.84      7165
           1       0.61      0.62      0.62      3046

    accuracy                           0.77     10211
   macro avg       0.73      0.73      0.73     10211
weighted avg       0.77      0.77      0.77     10211
```

Fig. 2. Classification report for DT+.

```
Accuracy of the model:
0.8225443149544609
Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.97      0.88      7165
           1       0.86      0.48      0.62      3046

    accuracy                           0.82     10211
   macro avg       0.84      0.73      0.75     10211
weighted avg       0.83      0.82      0.81     10211
```

Fig. 3. Classification report for RF.

```
Accuracy of the model:
0.823719518166683
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.95      0.88      7165
           1       0.81      0.53      0.64      3046

    accuracy                           0.82     10211
   macro avg       0.82      0.74      0.76     10211
weighted avg       0.82      0.82      0.81     10211
```

Fig. 4. Classification report for XGB.

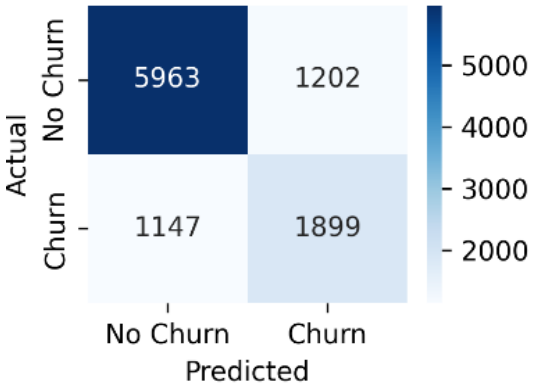

Fig. 5. Confusion matrix for DT+.

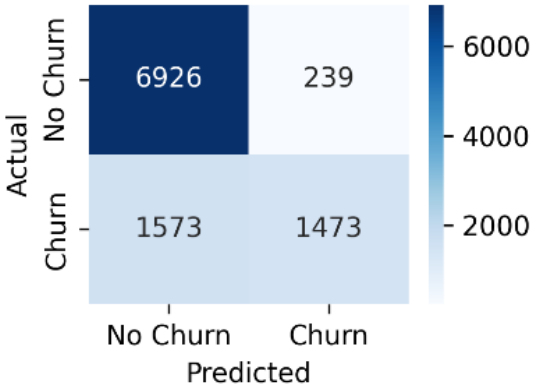

Fig. 6. Confusion matrix for RF.

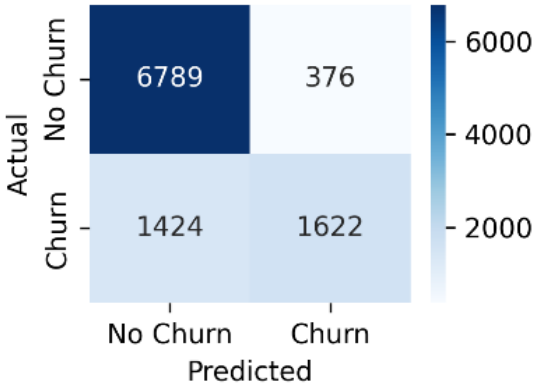

Fig. 7. Confusion matrix for XGB.

While the DT and DT+ models offered a balanced trade-off between recall and precision, their performance was constrained by inherent algorithmic limitations, such as susceptibility to overfitting. Grid search optimization for DT+ improved consistency but did not elevate its metrics significantly above those of the traditional DT model. The imputation and normalization steps proved vital, as they contributed to the balanced recall values across all models, ensuring accurate identification of churners. However, the reliance on advanced optimization techniques, such as those employed in XGBoost, demonstrates the need for robust preprocessing and model design to maximize predictive accuracy. Future work could explore alternative preprocessing strategies or combine these methods with deep learning models to achieve further improvements.

VII. CCP Performance Comparison with Existing CCP Methods Using The CELL2CELL Dataset

The Cell2Cell dataset has been extensively used in the telecommunications sector to develop and evaluate customer churn prediction (CCP) models. Various studies have explored various approaches, from traditional machine learning techniques to advanced deep learning frameworks, achieving diverse outcomes based on the methodologies and preprocessing techniques employed. The study in [26] applied a Deep-BP-ANN model combined with Lasso Regression and Variance Thresholding, achieving an accuracy of 79.38%. Their results outperformed traditional XGBoost and Logistic Regression models, demonstrating the potential of deep learning for churn prediction in imbalanced datasets. The study in [8] developed a hybrid ensemble approach combining clustering and classification algorithms, including k-medoids, Gradient-Boosted Trees (GBT), Decision Trees (DT), and Deep Learning (DL). This method achieved an accuracy of 93.6%, highlighting the effectiveness of integrating clustering techniques to manage class imbalance. The study in [19] explored decision forest models enhanced with weighted soft voting. Their approach achieved an accuracy of 96.57%, showcasing the advantages of ensemble methods in improving prediction accuracy and robustness by effectively identifying churn patterns.

Table II summarizes the performance metrics reported in these studies, providing a benchmark for assessing the efficacy of different CCP methodologies.

TABLE II. Performance Metrics Comparison from Existing Studies

Method	Accuracy	Precision	Recall	F1-Score	AUC
Deep-BP-ANN [26]	79.38	74.50	89.32	81.24	79.38
Hybrid.Ensemble [8]	93.6	79.10	67.45	72.81	93.6
Decision Forest [19]	96.57	96.57	85.45	83.72	96.57

Our study, which uses the Cell2Cell dataset with advanced preprocessing techniques, aligns with findings from these prior works while offering unique contributions. Preprocessing steps like KNN imputation, normalization, and SMOTE Tomek resampling effectively addressed class imbalance, enhancing model performance. Ensemble models such as RF and XGBoost achieved high accuracy (0.82), as highlighted in Table I, which compares the metrics of the models used in this study and key observations are as follows:

- Performance Context: Our study's ensemble approaches (RF and XGBoost) achieved competitive accuracy scores compared to more straightforward machine learning frameworks like DT and DT+. While slightly below the performance of Decision Forest reported by [19], the balance between interpretability and accuracy makes these models practical for real-world use.

- Preprocessing Impact: SMOTE Tomek and normalization proved essential in balancing class distribution, improving recall and precision across models. This approach parallels the findings of Liu et al. [8], who leveraged clustering techniques for similar benefits.
- Scalability and Practicality: Unlike computationally intensive deep learning models, XGBoost and RF offer a cost-effective solution for telecom companies. Their high accuracy and efficient preprocessing make them suitable for deployment in dynamic environments where real-time churn prediction is critical.

VIII. Real World Significance

The findings of this study have critical implications for addressing real-world challenges in the telecommunications industry, where customer churn remains a significant concern. By leveraging predictive models like XGBoost, telecom companies can transition from reactive to proactive churn management strategies, leading to tangible business benefits.

Customer churn directly impacts revenue and operational efficiency in saturated markets where acquiring new customers is significantly costlier than retaining existing ones. The predictive models evaluated in this study, particularly XGBoost and RF, provide robust tools for identifying high-risk customers. These insights empower telecom companies to design targeted retention strategies, such as personalized offers, improved customer service, or loyalty programs, mitigating churn effectively.

The predictive algorithms demonstrated in this study can be seamlessly integrated into Customer Relationship Management (CRM) systems. For instance, by embedding XGBoost into customer analytics platforms, companies can automate churn predictions and deliver actionable insights in real-time. The models' ability to handle high-dimensional data and imbalanced classes also ensures scalability, making them suitable for large, complex telecom datasets.

Proactive churn management driven by these models could result in measurable outcomes, including:

- Cost Savings: Reducing churn by even a small percentage can save millions in customer acquisition costs.
- Revenue Enhancement: Retaining high-value customers boosts recurring revenue and long-term profitability.
- Operational Efficiency: Automating churn prediction reduces manual analysis, freeing resources for strategic initiatives.

The study also addresses broader industry challenges, such as fostering customer loyalty in a competitive landscape. Predictive models not only assist in retaining existing customers but also enhance customer experience by anticipating needs and preferences. These advancements align with the strategic goals of telecom firms to sustain growth and remain competitive. The practical relevance of this research extends beyond theoretical improvements, offering telecom companies a pathway to adopt data-driven strategies for churn management. By implementing these models, businesses can achieve financial benefits and strengthen customer relationships, driving sustainable growth in a dynamic market environment.

IX. Conclusion and Future Work

This study comprehensively analyzed customer churn prediction (CCP) using a Decision Tree (DT). The Decision Tree improved with grid search (DT+), Random Forest (RF), and XGBoost (XGB) algorithms applied to the preprocessed Cell2Cell dataset. Among these models, XGBoost emerged as the most effective, achieving strong precision, recall, and F1 scores with an accuracy of 0.82. Its advanced optimization techniques, such as gradient boosting and error correction, maximized the benefits of preprocessing methods, demonstrating its superiority in handling complex and high-dimensional datasets. The RF model also delivered robust performance, achieving an accuracy of 0.82 while effectively balancing precision and recall. Its ensemble nature successfully mitigated overfitting and enhanced generalization. In contrast, the DT+ model, despite its improvements through grid search, faced limitations in reaching comparable performance, underscoring the inherent constraints of decision tree-based models.

This research contributes to the field by integrating robust preprocessing techniques, such as KNN imputation, normalization, and SMOTE Tomek resampling, with grid search optimization. These methodologies collectively address critical challenges posed by imbalanced and high-dimensional telecom datasets, providing a scalable and systematic framework for CCP. By evaluating the comparative performance of DT, RF, and XGBoost models, the study underscores the value of ensemble methods and advanced hyperparameter tuning in achieving accurate and reliable predictions. Furthermore, the findings validate the significance of preprocessing as a cornerstone for effective churn prediction, offering insights into how these techniques enhance model robustness.

While this study provides valuable insights, certain limitations should be acknowledged. The analysis is constrained to the Cell2Cell dataset, which, while comprehensive, may not fully represent the diversity of customer behaviors in other industries or regions. Additionally, the models used in this study rely heavily on preprocessing and hyperparameter tuning, which may increase computational costs for large-scale datasets. Although effective, the interpretability of ensemble methods like XGBoost can be challenging, potentially limiting their application in contexts requiring high transparency.

Future research could explore incorporating deep learning models like neural networks that integrate different classifiers for enhanced predictive power. Investigating advanced feature engineering techniques, such as automated feature selection and interaction effects, could further improve model performance. Additionally, extending this research to other industries, such as finance or retail, with diverse customer behaviors would help validate the generalizability of the proposed methods. Exploring real-time churn prediction systems and using external data sources, such as social media

or customer feedback, could also offer new avenues for development.

The study's findings hold significant practical implications for the telecommunications industry. By providing actionable insights into the design and optimization of predictive models, this research supports proactive customer retention strategies, enabling telecom companies to reduce churn rates and enhance profitability. Moreover, the methodologies presented in this work contribute to advancing the knowledge base in CCP, offering scalable and interpretable solutions for addressing the challenges of imbalanced and high-dimensional datasets.

ACKNOWLEDGMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editor and the reviewers. Any statements, claims, performances and results are not guaranteed or endorsed by the publisher.

REFERENCES

[1] H. K. Thakkar, A. Desai, S. Ghosh, P. Singh, and G. Sharma, "Clairvoyant: AdaBoost with Cost-Enabled Cost-Sensitive Classifier for Customer Churn Prediction," Comput. Intell. Neurosci., vol. 2022, no. 1, p. 9028580, 2022, doi: 10.1155/2022/9028580.

[2] T. Zhang, S. Moro, and R. F. Ramos, "A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation," Future Internet, vol. 14, no. 3, Art. no. 3, Mar. 2022, doi: 10.3390/fi14030094.

[3] A. Amin, A. Adnan, and S. Anwar, "An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes," Appl. Soft Comput., vol. 137, p. 110103, Apr. 2023, doi: 10.1016/j.asoc.2023.110103.

[4] A. Khattak, Z. Mehak, H. Ahmad, M. U. Asghar, M. Z. Asghar, and A. Khan, "Customer churn prediction using composite deep learning technique," Sci. Rep., vol. 13, no. 1, p. 17294, Oct. 2023, doi: 10.1038/s41598-023-44396-w.

[5] S. O. Abdulsalam, M. O. Arowolo, Y. K. Saheed, and J. O. Afolayan, "Customer Churn Prediction in Telecommunication Industry Using Classification and Regression Trees and Artificial Neural Network Algorithms," Indones. J. Electr. Eng. Inform. IJEEI, vol. 10, no. 2, Art. no. 2, Jun. 2022, doi: 10.52549/ijeei.v10i2.2985.

[6] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," Comput. Math. Organ. Theory, vol. 25, pp. 319–335, 2019.

[7] K. Cabello-Solorzano, I. Ortigosa de Araujo, M. Peña, L. Correia, and A. J. Tallón-Ballesteros, "The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis," in 18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023), P. García Bringas, H. Pérez García, F. J. Martínez de Pisón, F. Martínez Álvarez, A. Troncoso Lora, Á. Herrero, J. L. Calvo Rolle, H. Quintián, and E. Corchado, Eds., Cham: Springer Nature Switzerland, 2023, pp. 344–353. doi: 10.1007/978-3-031-42536-3_33.

[8] R. Liu et al., "An Intelligent Hybrid Scheme for Customer Churn Prediction Integrating Clustering and Classification Algorithms," Appl. Sci., vol. 12, no. 18, Art. no. 18, Jan. 2022, doi: 10.3390/app12189355.

[9] N. Edwine, W. Wang, W. Song, and D. Ssebuggwawo, "Detecting the Risk of Customer Churn in Telecom Sector: A Comparative Study," Math. Probl. Eng., vol. 2022, p. e8534739, Jul. 2022, doi: 10.1155/2022/8534739.

[10] G. Jiao and H. Xu, "Analysis and Comparison of Forecasting Algorithms for Telecom Customer Churn," J. Phys. Conf. Ser., vol. 1881, no. 3, p. 032061, Apr. 2021, doi: 10.1088/1742-6596/1881/3/032061.

[11] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, "Customer churn prediction in telecom sector using machine learning techniques," Results Control Optim., vol. 14, p. 100342, Mar. 2024, doi: 10.1016/j.rico.2023.100342.

[12] S. O. Abdulsalam, J. F. Ajao, B. F. Balogun, and M. O. Arowolo, "A Churn Prediction System for Telecommunication Company Using Random Forest and Convolution Neural Network Algorithms," EAI Endorsed Trans. Mob. Commun. Appl., vol. 7, no. 21, Jul. 2022, Accessed: Mar. 30, 2024. https://eudl.eu/doi/10.4108/eetmca.v6i21.2181

[13] L. F. Khalid, A. Mohsin Abdulazeez, D. Q. Zeebaree, F. Y. H. Ahmed, and D. A. Zebari, "Customer Churn Prediction in Telecommunications Industry Based on Data Mining," in 2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA), Jul. 2021, pp. 1–6. doi: 10.1109/ISIEA51897.2021.9509988.

[14] H. Jain, A. Khunteta, and S. P. Shrivastav, "Telecom churn prediction using seven machine learning experiments integrating features engineering and normalization," 2021, Accessed: Apr. 08, 2024. https://www.researchsquare.com/article/rs-239201/latest

[15] D. D. Adhikary and D. Gupta, "Applying over 100 classifiers for churn prediction in telecom companies," Multimed. Tools Appl., vol. 80, no. 28, pp. 35123–35144, Nov. 2021, doi: 10.1007/s11042-020-09658-z.

[16] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," Computing, vol. 104, no. 2, pp. 271–294, Feb. 2022, doi: 10.1007/s00607-021-00908-y.

[17] R. P. Sari, F. Febriyanto, and A. C. Adi, "Analysis Implementation of the Ensemble Algorithm in Predicting Customer Churn in Telco Data: A Comparative Study," Informatica, vol. 47, no. 7, Art. no. 7, Jul. 2023, doi: 10.31449/inf.v47i7.4797.

[18] M. Imani, Z. Ghaderpour, and M. Joudaki, "The Impact of SMOTE and ADASYN on Random Forests and Advanced Gradient Boosting Techniques in Telecom Customer Churn Prediction," Mar. 05, 2024, Preprints: 2024030213. doi: 10.20944/preprints202403.0213.v1.

[19] F. E. Usman-Hamza et al., "Intelligent Decision Forest Models for Customer Churn Prediction," Appl. Sci., vol. 12, no. 16, Art. no. 16, Jan. 2022, doi: 10.3390/app12168270.

[20] M. E. Saleh, N. Abd-Alsabour "On the impact of various combinations of preprocessing steps on customer churn prediction," unpublished.

[21] B. Ramosaj and M. Pauly, "Predicting missing values: a comparative study on non-parametric approaches for imputation," Comput. Stat., vol. 34, no. 4, pp. 1741–1764, Dec. 2019, doi: 10.1007/s00180-019-00900-3.

[22] Y. Fareniuk, T. Zatonatska, O. Dluhopolskyi, and O. Kovalenko, "Customer churn prediction model: a case of the telecommunication market," ECONOMICS, vol. 10, no. 2, pp. 109–130, Dec. 2022.

[23] W. H. Khoh, Y. H. Pang, S. Y. Ooi, L.-Y.-K. Wang, and Q. W. Poh, "Predictive Churn Modeling for Sustainable Business in the Telecommunication Industry: Optimized Weighted Ensemble Machine Learning," Sustainability, vol. 15, no. 11, Art. no. 11, Jan. 2023, doi: 10.3390/su15118631.

[24] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," Appl. Soft Comput., vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.

[25] N. N. Y, T. V. Ly, and D. V. T. Son, "Churn prediction in telecommunication industry using kernel Support Vector Machines," PLOS ONE, vol. 17, no. 5, p. e0267935, May 2022, doi: 10.1371/journal.pone.0267935.

[26] S. Wael Fujo, S. Subramanian, and M. A. Khder, "Customer Churn Prediction in Telecommunication Industry Using Deep Learning," Inf. Sci. Lett., vol. 11, no. 1, Dec. 2021, [Online]. Available: https://digitalcommons.aaru.edu.jo/isl/vol11/iss1/24