KJCS
Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 221-232
https://doi.org//10.47514/kjcs/2026.3.3.005

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

An Explainable Random Forest Framework with Feature Engineering for Customer Churn in Telecommunications

Bappa Mohammad Gambo[1], Murtala Muhammad[2], Zanna Bulama[3] and Adamu Liman Mohammed[4]
[1,2,3,4] Department of Computer Science, Modibbo Adama University, Yola, Nigeria
gambobappa1@gmail.com[1]

Abstract

Customer churn is a major problem for the telecommunications industry because it results in loss of revenue and increase in customer acquisition costs. Existing churn prediction methods mainly focus on the predictive accuracy and lack of explanations for the predictions. The study proposed an intelligent customer churn prediction framework combining predictive modeling and explainable artificial intelligence techniques to facilitate customer retention decision making. The objectives were to preprocess and prepare the customer data, engineer informative features, tackle class imbalance, develop and simulate a predictive model and provide explainable churn predictions. The dataset used in the study was Telco Customer Churn Dataset from Kaggle. Data preprocessing, feature engineering, feature scaling and class balancing using Synthetic Minority Oversampling Technique (SMOTE) were applied. A Random Forest classifier was trained as predictive engine and used feature importance analysis and SHapley Additive exPlanations (SHAP) to provide global and local explanations of model predictions. The developed model attained the accuracy of 79% with the precision of 0.6148, recall of 0.6192, F1-score of 0.6170, and receiver operating characteristic area under the curve (ROC-AUC) score of 0.8473. The results of the simulation showed practical value of the framework in finding customers that are likely to churn. The study concluded that the proposed framework is effective in predicting customer churn, while at the same time providing interpretable predictions. Telecommunication companies should embrace explainable predictive systems for proactive customer retention strategies.

Keywords*:* customer churn prediction, random forest, explainable artificial intelligence, SHAP, feature engineering, SMOTE, telecommunication, machine learning.

1. Introduction

Customer retention has become a critical issue in the telecommunications industry with the increasing competition in the telecom market and high costs of acquiring new customers(Nurtriana et al., 2024). Customer churn is the loss of existing subscribers due to discontinuing service, and can have a major impact on an organization's revenue, profitability and long-term sustainability. Consequently, telecom companies are increasingly turning to data-driven techniques to identify customers who are likely to churn and to implement proactive retention measures(Asif et al., 2025). Machine learning techniques have become powerful tools for predicting customer churn due to their ability to process extensive customer data and identify complex behavioral patterns. Among these, Random Forest has been widely adopted because of its robustness, ability to handle heterogeneous data and its predictive performance(Boozary et al., 2025). However, most of the existing churn prediction models are constructed mainly based on the prediction accuracy with poor

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 221-232
https://doi.org//10.47514/kjcs/2026.3.3.005
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

understanding of the reasons behind the prediction. This lack of transparency can reduce stakeholder trust and hinder the practical adoption of predictive systems in business environments(Bhargav Bhushan, 2025). Furthermore, customer churn datasets typically have problems like class imbalance and missing feature representation. Class imbalance arises when there is a significant difference in the number of non-churners versus churners, leading to biased predictions for the majority class (Suguna et al., 2025). Similarly, raw customer data may not be sufficient to represent behavioral patterns relevant to churn prediction without suitable feature engineering techniques (Chang et al., 2024). Recent advances in Explainable Artificial Intelligence (XAI) have led to mechanisms to improve the interpretability of machine learning models. Techniques like feature importance analysis and SHapley Additive exPlanations (SHAP) helps stakeholders understand both global model behavior and the results of individual prediction outcomes(Xu et al., 2023). However, there are only a few studies that integrate feature engineering, class balancing, predictive modelling, simulation and explainability into a single framework for customer churn prediction. To address this issue, this study proposes an Explainable Random Forest Framework with Feature Engineering for customer churn in telecommunications. The framework comprises data preprocessing, feature engineering, class balancing using Synthetic Minority Oversampling Technique (SMOTE), Random Forest classification, model simulation and explainable artificial intelligence techniques. The goal is to not only enhance the predictive performance but also to deliver transparent and interpretable predictions which can support customer retention decision making.

2. Related Works

Customer churn prediction has become a hot topic lately, and for good reason it is crucial for keeping customers and managing revenue effectively. A number of studies have explored how machine learning can be used to predict churn, especially in the telecommunication industry. For instance, Fujo et al. (2022) looked into how class imbalance affects churn prediction and found that using the Synthetic Minority Oversampling Technique (SMOTE) alongside ensemble classifiers significantly boosted recall and F1-score. In a similar vein, Asif et al. (2025) discovered that combining SMOTE with ensemble models like Random Forest and Gradient Boosting improved churn detection while still keeping overall classification accuracy intact. These findings really emphasize the need to tackle class imbalance when creating churn prediction models. Ensemble learning methods have also gained traction in churn prediction. Usman-Hamza et al. (2022) introduced an Intelligent Decision Forest framework that surpassed traditional classifiers in both predictive accuracy and robustness. Meanwhile, Nurtriana et al. (2024) compared Logistic Regression, Support Vector Machines, and Random Forest models, finding that Random Forest outperformed the others. Chang et al. (2024) took it a step further, showing that ensemble learning frameworks consistently delivered better accuracy and recall compared to individual classifiers. All of this points to the fact that tree-based ensemble methods are particularly adept at capturing the intricate patterns of customer behavior that lead to churn. Recent research has been increasingly weaving explainable artificial intelligence (XAI) techniques into churn prediction systems. Maan & Maan (2023) took a step forward by integrating SHAP with ensemble learning models, which

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

led to better transparency and feature selection in their models. Poudel et al. (2024) highlighted how crucial explainability is in churn prediction, showing that SHAP effectively pinpointed key factors driving churn, such as contract type, tenure, and monthly charges. In a similar vein, SHAP's effectiveness has been demonstrated through its ability to provide both global and local insights for telecom churn predictions, which in turn bolstered stakeholder trust and enhanced the interpretability of the results (Bhushan, 2025). Feature engineering has also emerged as a vital element in boosting predictive performance. Feature engineering is the process of transforming raw data into meaningful representations that enhance predictive performance (Fashanu et al., 2025). Xu et al. (2023) showed that a business focused approach to feature engineering significantly improved the accuracy and interpretability of churn predictions. Additionally, Boozary et al. (2025) found that merging engineered features with adaptive ensemble learning strategies enhanced the detection of churn patterns and overall predictive performance. However, despite these great strides in predicting customer churn, most research has tackled aspects like feature engineering, handling class imbalances, predictive modeling, and explainability as separate issues. As a result, there has been a lack of focus on creating a cohesive framework that brings together data preprocessing, feature engineering, class balancing, predictive modelling, and explainable AI all in one churn prediction pipeline. Additionally, many current methods prioritize predictive accuracy but fall short in offering the transparency and actionable insights needed for effective customer retention strategies. This study aims to develop a cohesive framework that comprises data prev processing, feature engineering, class balancing, predictive modeling, and explainable AI in one pipeline to provide meaningful and actionable information for customer retention strategies in the telecommunication industry by introducing an Explainable Random Forest Framework with Feature Engineering for Customer Churn in Telecommunications.

3. Methodology

In this study, the IBM Telco Customer Churn dataset from Kaggle was used, which is a widely used benchmark dataset for customer churn prediction research. The dataset consists of information for 7,043 telecommunications customers, including demographic data, account data, billing data and service usage characteristics. The target variable is the attribute Churn. This attribute shows if the customer has left the service (Yes) or if the customer is still subscribed (No).

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 221-232
https://doi.org//10.47514/kjcs/2026.3.3.005
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

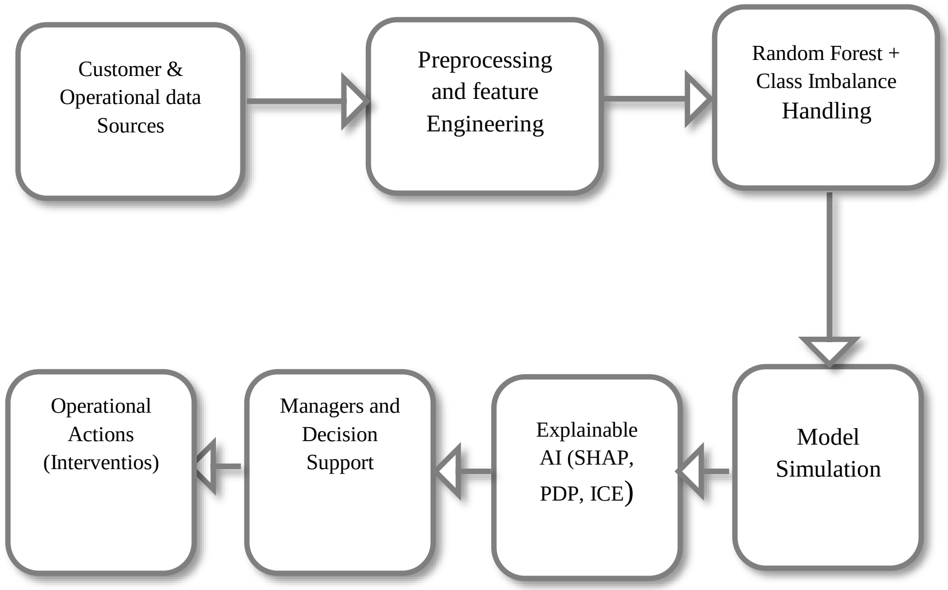

Figure 1: Proposed System Model

Figure 1 shows the proposed explainable framework for customer churn prediction. The framework starts with data acquisition from IBM Telco Customer Churn dataset, followed by data pre-processing and feature engineering. The processed dataset was divided into three subsets: training, validation, and testing data. 70% of the dataset was assigned to the training set, and 15% was used for validation and the remaining 15% was kept for testing. The random forest model was trained on the training data; the validation data were used during model development for hyperparameter selection and the testing data were kept separate for the final evaluation of the trained model. The data were partitioned using stratified sampling in an effort to maintain the representative distribution of the churn and non-churn classes across the three subsets. Then smote is applied to the training data to overcome class imbalance. Then, Random Forest classifier will be train to predict customer churn. In this work, model simulation is use to evaluate the predictions and feature importance analysis and SHAP explanations to interpret them. The insights produced can help to guide decision making on customer retention and churn management strategies.

3.1 Dataset Preprocessing and Feature Engineering

The dataset was pre-processed to enhance the quality of data and prepare it for machine learning analysis. The data cleaning process included dealing with missing data, removing irrelevant features, encoding categorical variables into numerical format, and normalizing numerical features. Feature engineering was done to improve the predictive power of the model by creating new features from the existing ones.

Table 1. Engineered features

Original Feature	Engineered Feature
Total Charges	Avg_Monthly_Spend
Tenure Months	Tenure_Group
Monthly Charges	High_Value
Service Features	Total_Services

These engineered features were intended to capture customer behavioral patterns that were not explicitly represented in the original dataset. The dataset was divided into training and testing subsets before feature scaling and model training were performed in other to prevent biased performance evaluation and prevent data leakage.

3.2 Class Balancing and Random Forest Classification

Customer churn datasets are often imbalanced, meaning that non churning customers outnumber churning customers by a large margin. To tackle this issue, Synthetic Minority Oversampling Technique (SMOTE) was applied to the training dataset to create synthetic samples for the minority churn class. The processed data set was then used to train a Random Forest classifier following class balancing. Random Forest was chosen for its robustness, ability to model complex non linear relationships, resistance to overfitting and feature importance measures for interpreting the models.

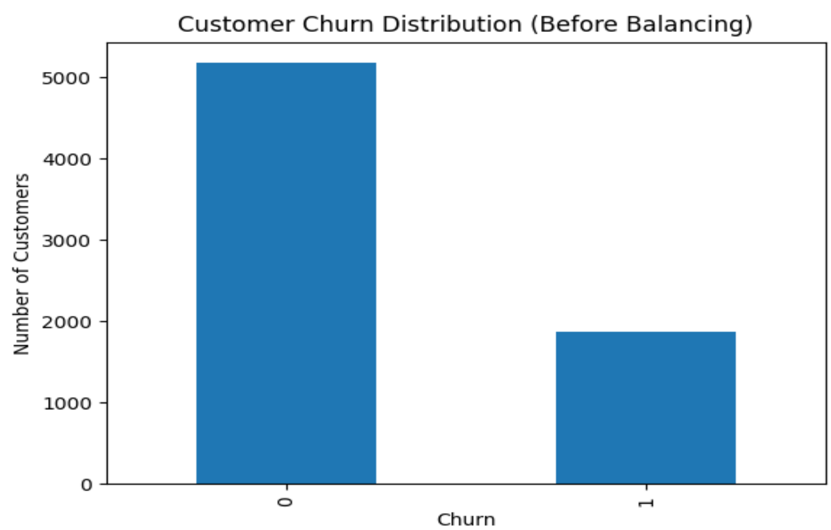

Figure 2: IBM telco dataset class distribution

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Figure 2 shows the class distribution of the dataset before the application of SMOTE. It clearly shows that the non-churning customers significantly outnumber the churning ones, which if not balanced, will skew the machine learning toward the majority class.

3.3 Validation and Hyperparameter Tuning

The dataset was partitioned into training, validation and test subsets. A validation set was employed for assisting model development and hyperparameter selection. During the validation stage, a suitable configuration of the Random Forest was identified before the final model was evaluated on unseen testing data. This procedure provided an additional evaluation stage between model training and final testing and helped to reduce the possibility of selecting model parameters based on the test data.

3.4 Model Simulation and Explainability

To replicate real world churn prediction scenarios, the trained model was evaluated on unseen customer records in the testing dataset. The simulation offered individual customer churn classifications and churn probabilities, thus proving the practical applicability of the framework. Transparency and interpretability were improved by applying explainable artificial intelligence techniques. Feature importance analysis was used for global model behavior analysis. Local prediction explanations were generated using SHapley Additive exPlanations (SHAP), which allowed to identify factors that influence individual churn predictions.

3.5 Evaluation Metrics

The performance of the proposed framework was evaluated in terms of Accuracy, Precision, Recall, F1-score and Receiver Operating Characteristic Area Under the Curve (ROC-AUC).

Metrics	Purpose
Accuracy	Measures the overall proportion of correctly classified customers
Precision	Measures the proportion of customers predicted as churners who actually churned
Recall	Measures the proportion of actual churners correctly identified by the model
F1-score	Provides a balanced measure of Precision and Recall.
ROC-AUC	Measures the model's ability to distinguish between churners and non-churners across different classification thresholds

Table 2. Evaluation Metrics Used for Model Assessment

Accuracy evaluated the general correctness of predictions, while Precision, Recall and F1-score evaluated the effectiveness of the model in finding churners. The ROC-AUC was used to assess the ability of the model to differentiate between churning and non-churning customers at various classification thresholds.

4. Results and Discussions

The original dataset showed a significant imbalance between customers who churn and those who don’t. In figure 3, class distribution after the application of the synthetic minority oversampling technique was depicted. illustrating the change in class representation following oversampling.

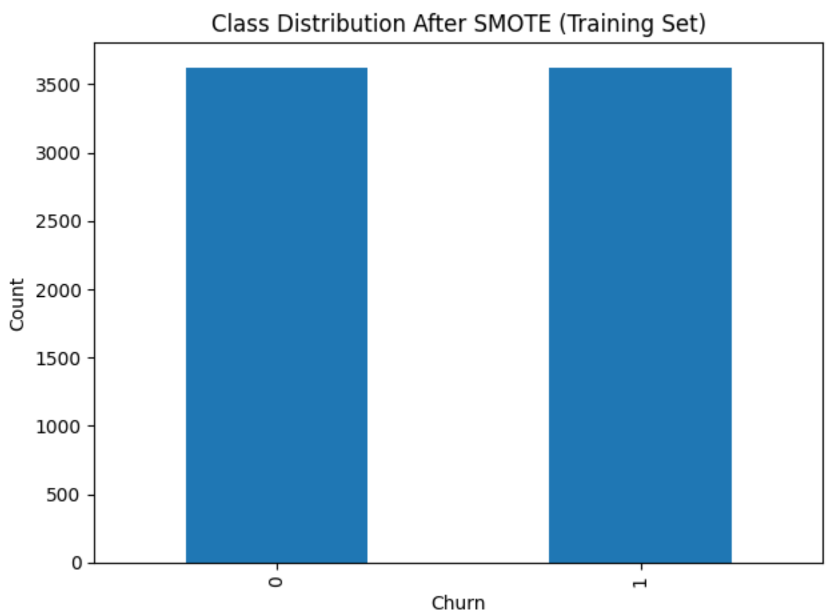

Figure 3: Class Distribution After SMOTE

Initially, the dataset showed a significant class imbalance, with non churning customers vastly outnumbering those who churn. This kind of imbalance can skew machine learning models towards the majority class, making it harder for them to spot customers who might be at risk of churning. After implementing SMOTE, the minority churn class was balanced by creating synthetic samples, which led to a better representation of classes during model training.

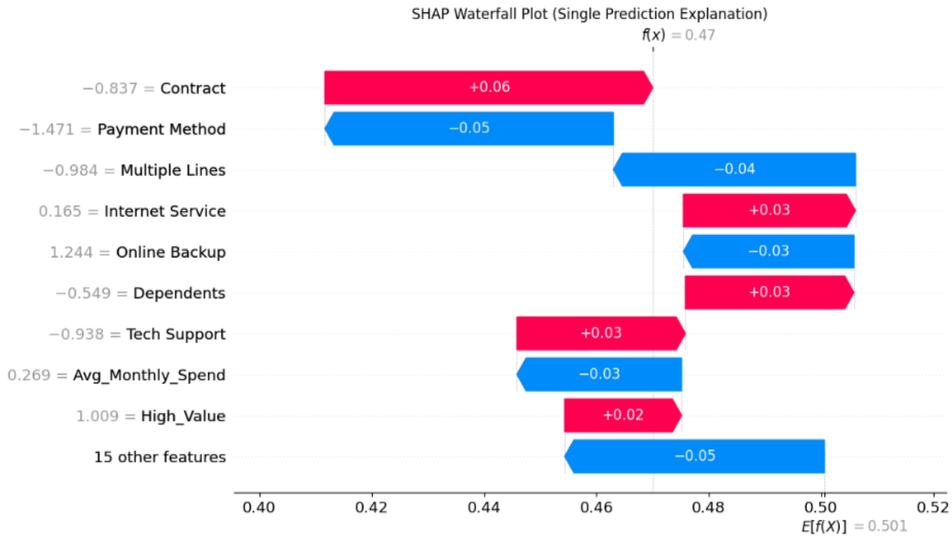

Figure 4: Feature Importance

In Figure 4, the feature importance analysis sheds light on how the model behaves overall. It reveals that factors like Contract, Tenure Months, Monthly Charges, Total Charges, Average Monthly Spend, Tenure Group, and Customer Lifetime Value (CLTV) play significant roles in predicting churn. This suggests that elements like customer commitment, how long they've been with the service, their spending habits, and their overall value as customers are key factors in churn within the telecommunications industry.

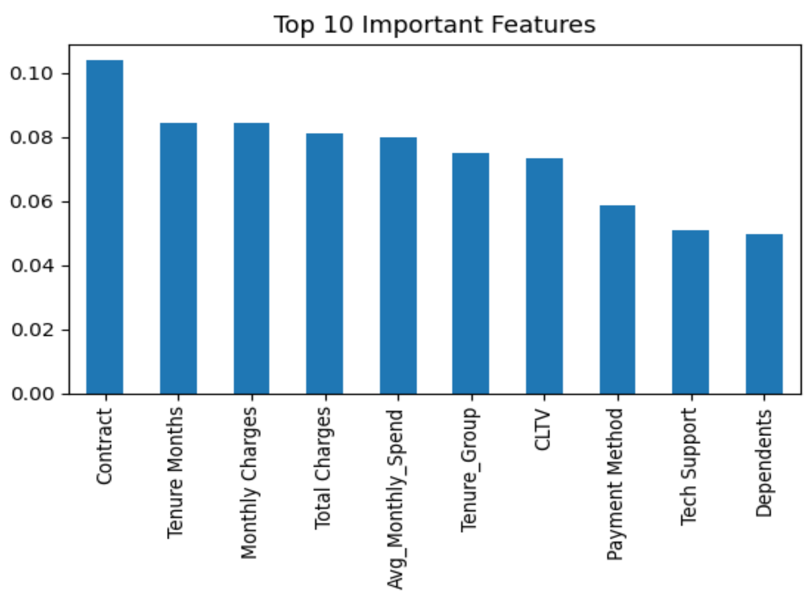

Figure 5: SHAP Waterfall Plot

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 221-232
https://doi.org//10.47514/kjcs/2026.3.3.005
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

To enhance model transparency, SHAP was used to provide local explanations for individual predictions. The SHAP waterfall plot in Figure 5 shows how specific customer characteristics influenced the predicted churn outcome. Attributes like Contract, Internet Service, Dependents, and Tech Support were linked to a higher churn probability, while factors such as Payment Method, Multiple Lines, and Average Monthly Spend helped lower the churn risk. This SHAP analysis illustrates that the proposed framework not only delivers accurate predictions but also offers clear explanations that can aid in managerial decision making and strategies for retaining customers.

4.1 Model Evaluation

The predictive performance of the Random Forest model was examined using the confusion matrix, which provides a detailed view of the model's classification outcomes. Figure 6 presents the confusion matrix, showing the numbers of correctly and incorrectly classified churn and non churn customers.

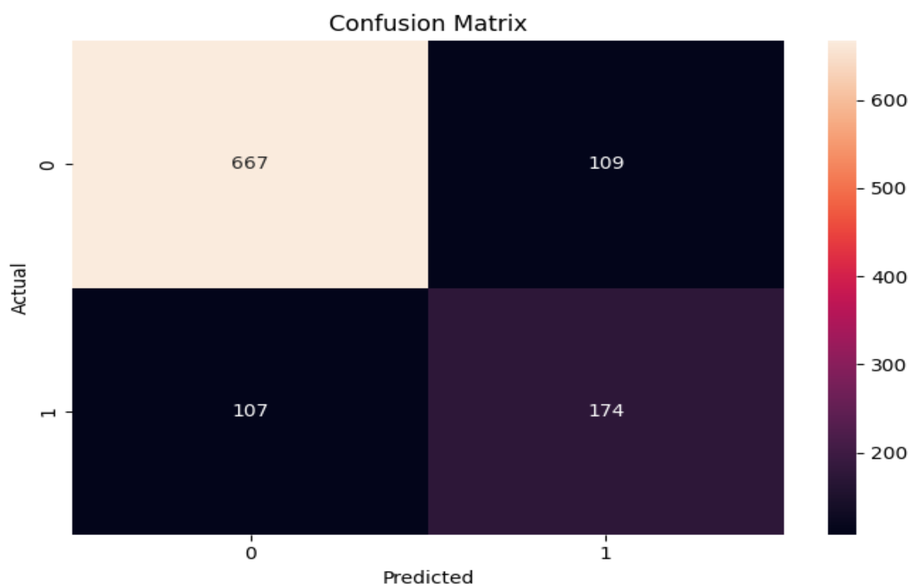

Figure 6: Confusion Matrix

The confusion matrix reveals that the model successfully classified 667 non-churning customers and 174 churning customers. There were some misclassifications, including 109 false positives and 107 false negatives. Overall, these results shows that the model effectively differentiated between churners and non-churners with a commendable level of accuracy, while still being reasonably sensitive to churn cases.

4.2 Performance metrics

The overall predictive performance of the Random Forest model was further assessed using Accuracy, Precision, Recall, F1-score, and ROC-AUC. Table 3 showed the performance results

Kasu Journal of Computer Science Vol.3 No.3 [June, 2026], pp. 221-232
https://doi.org//10.47514/kjcs/2026.3.3.005
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

obtained on the test dataset, giving a complete assessment of the model's classification effectiveness and its ability to distinguish between churning and non-churning customers.

Table 3. Classification Report

	Precision	**Recall**	**F1-Score**	**support**
0	0.8618	0.8595	0.8606	776
1	0.6148	0.6192	0.6170	281
Accuracy			0.7956	1057
Macro Avg	0.7383	0.7394	0.7388	1057
Weighted Avg	0.7961	0.7956	0.7959	1057

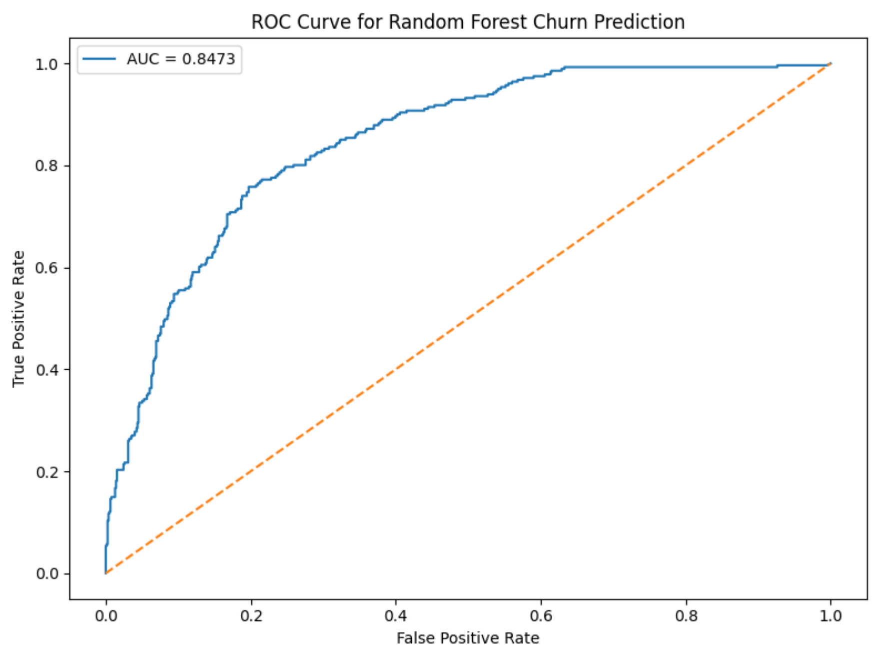

Figure 7: ROC-AUC Curve

The classification results show that the Random Forest classifier hit an overall accuracy of 79%. For churn prediction, the model achieved a precision of 0.6148, a recall of 0.6192, and an F1-score of 0.6170. These metrics reflect a well-balanced predictive performance, highlighting the model's capability to spot customers who might leave the service. Even more impressive, the proposed framework scored a ROC-AUC of 0.8473, showcasing its strong ability to distinguish between

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

churning and non churning customers. The ROC curve in Figure 7 backs this up, as it stays well above the diagonal reference line that represents random classification.

4.3 Major Findings

The key findings from the experimental results are summarized below. These insights showcased how well the proposed framework predicts outcomes, identify the main factors linked to customer churn, and emphasize the role of explainability in understanding the model's predictions.

1. By combining feature engineering with SMOTE based class balancing, how customer behavior is represented is significantly improved and class imbalance issue is tackled. This enhancement boosted the model's ability to spot potential churners.
2. The proposed Random Forest framework delivered impressive predictive performance, achieving an accuracy of 79% and a ROC-AUC score of 0.8473. This shows its strong ability to differentiate between customers who are likely to churn and those who are not.
3. Analyzing feature importance highlighted that factors like Contract type, Tenure Months, Monthly Charges, Total Charges, Average Monthly Spend, and Customer Lifetime Value (CLTV) play a crucial role in influencing customer churn within the telecommunications industry.
4. By integrating explainable artificial intelligence techniques using SHAP, clear explanations were provided for individual predictions. This approach not only ensures accurate churn predictions but also offers interpretable decision support for managing customer retention effectively.

5. Conclusion

This study introduced an Explainable Random Forest Framework tailored for tackling Customer Churn in Telecommunications. It does this by seamlessly combining data preprocessing, feature engineering, SMOTE based class balancing, Random Forest classification, model simulation, and explainable AI techniques. The framework showed impressive predictive performance, achieving an accuracy of 79% and a ROC-AUC score of 0.8473. Plus, it offers both global and local insights into model predictions through feature importance analysis and SHAP. These results highlight the framework's ability to provide accurate and interpretable churn predictions, making it an essential tool for decision making in customer retention management within the telecommunications industry. Based on the results of this study, Telecommunication companies are highly recommended to embrace predictive analytics frameworks for managing customer churn. This way, they can spot customers who might be thinking about leaving and take proactive steps to keep them before they actually do. Moreover, future research directions include the application of advanced machine learning and deep learning techniques such as Gradient Boosting, XGBoost, LightGBM and Artificial Neural Networks to see if higher values of precision, recall and F1-score can be obtained for customer churn prediction.

References

Asif, D., Arif, M. S., & Mukheimer, A. (2025). A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry. *Results in Engineering*, *26*. https://doi.org/10.1016/j.rineng.2025.104629

Bhargav Bhushan, S. (2025). *Enhancing Customer Churn Prediction in the Telecom Sector Using Advance Machine Learning Techniques and Explainable AI MSc Research Project MSc Data Analytics*.

Boozary, P., Sheykhan, S., GhorbanTanhaei, H., & Magazzino, C. (2025). Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction. *International Journal of Information Management Data Insights*, *5*(1). https://doi.org/10.1016/j.jjimei.2025.100331

Chang, V., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., & Benson, V. (2024). Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models. *Algorithms*, *17*(6). https://doi.org/10.3390/a17060231

Fashanu, J. O., Abdulmalik, U. I., & Kassim, M. (2025). An enhanced rice grain classification using a novel feature engineering and random forest modelling. KASU Journal of Computer Science, 2(1), 33–40. https://www.kjcs.com.ng/view_paper/92

Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). Customer churn prediction in telecommunication industry using deep learning. *Information Sciences Letters*, *11*(1), 185–198. https://doi.org/10.18576/isl/110120

Maan, J., & Maan, H. (2023). Customer Churn Prediction Model using Explainable Machine learning. *International Journal of Computer Science Trends and Technology*, *11*. Retrieved www.ijcstjournal.org

Nurtriana, A., Rachmawati, D. D., Artiyasa, M., & Sidiq, D. S. Z. (2024). Churn prediction analysis of telecom customers using svm, random forest and logistic regression models using orange data mining tools. *E3S Web of Conferences*, *501*. https://doi.org/10.1051/e3sconf/202450102012

Özkurt, C. (2025). Comparative Analysis of XAI Techniques on Telecom Churn Prediction Using SHAP and Interpreted ML Partial Dependence. *Türk Doğa ve Fen Dergisi*, *14*(2), 11–25. https://doi.org/10.46810/tdfd.1529139

Poudel, S. S., Pokharel, S., & Timilsina, M. (2024). Explaining customer churn prediction in telecom industry using tabular machine learning models. *Machine Learning with Applications*, *17*, 100567. https://doi.org/10.1016/j.mlwa.2024.100567

Suguna, R., Suriya Prakash, J., Aditya Pai, H., Mahesh, T. R., Vinoth Kumar, V., & Yimer, T. E. (2025). Mitigating class imbalance in churn prediction with ensemble methods and SMOTE. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-01031-0

Usman-Hamza, F. E., Balogun, A. O., Capretz, L. F., Mojeed, H. A., Mahamad, S., Salihu, S. A., Akintola, A. G., Basri, S., Amosa, R. T., & Salahdeen, N. K. (2022). Intelligent Decision Forest Models for Customer Churn Prediction. *Applied Sciences (Switzerland)*, *12*(16). https://doi.org/10.3390/app12168270

Xu, T., Ma, Y., Ao, C., Qu, M., & Meng, X. H. (2023). A Novel Telecom Customer Churn Analysis System Based On Rfm Model and Feature Importance Ranking. *Interdisciplinary Journal of Information, Knowledge, and Management*, *18*, 719–737. https://doi.org/10.28945/5192