World Journal of Engineering and Technology
www.worldengineerjournal.com
Online ISSN:
Received: 18-07-2025, Accepted: 17-08-2025, Published: 02-09-2025
Volume 1, Issue 1, Jul-Sep, 2025, Page No. 7-11

Enhanced predictive data mining algorithm for fraud detection and churn modelling in telecommunication systems

Odudu Bassey Michael
Centre for Information and Communication Technology University of Port Harcourt, Nigeria

Abstract
The telecommunications industry operates in an intensely competitive and data-rich environment, where two of the most critical challenges are fraudulent activities and customer churn. These issues result in billions of dollars in annual revenue loss, making their effective mitigation a top priority for operational sustainability and profitability. Traditional data mining and statistical approaches often struggle with the high-dimensional, imbalanced, and complex nature of telecom data, leading to suboptimal detection rates and high false positives. This paper proposes an Enhanced Predictive Data Mining Algorithm (EPDMA) that integrates advanced feature engineering, a hybrid sampling technique to address class imbalance, and a stacked ensemble learning model to simultaneously and more accurately address both fraud detection and churn prediction. The feature engineering phase leverages Recursive Feature Elimination (RFE) and domain-specific feature creation, such as rolling window statistics and behaviour change metrics. To handle the severe class imbalance inherent in both problems, a combination of Synthetic Minority Over-sampling Technique (SMOTE) and Edited Nearest Neighbours (ENN) is employed. The core predictive model is a stacked ensemble comprising diverse base learners, including Gradient Boosting Machines (XGBoost), Random Forest, and a deep neural network, whose outputs are meta-learned by a logistic regression model for final prediction. The proposed EPDMA was rigorously evaluated on real-world, anonymized telecom datasets and benchmarked against state-of-the-art individual classifiers like XGBoost, Support Vector Machines (SVM), and traditional logistic regression. The results demonstrate a superior performance, with the EPDMA achieving an average precision of 94.5% and a recall of 89.2% for fraud detection, and an AUC-ROC score of 0.935 for churn modelling, significantly outperforming all baseline models. The discussion elucidates how the integration of these techniques overcomes the limitations of singular models, and the conclusion highlights the practical implications of deploying such a robust system for telecom operators, ultimately leading to enhanced security, improved customer retention strategies, and substantial financial savings.

Keywords: Telecommunications, fraud detection, churn modelling, predictive data mining, ensemble learning, machine learning, smote, stacking classifier, feature engineering, xgboost, customer retention, imbalanced data.

Introduction

The telecommunications sector stands as a cornerstone of the global digital economy, facilitating not only personal communication but also serving as the critical infrastructure for countless other industries and services. In this hyper-competitive landscape, characterized by saturated markets and thin profit margins, telecom operators face persistent and evolving threats that directly impact their revenue and operational integrity. Among these, fraudulent activities and customer churn represent two of the most formidable and financially draining challenges. Telecom fraud encompasses a wide array of illicit activities, including but not limited to Subscription Fraud, where individuals sign up for service with no intention of payment; International Revenue Share Fraud (IRSF), which artificially inflates traffic to high-premium numbers; and Wangiri or "one-ring" scams. The financial implications are staggering, with global losses estimated to be in the tens of billions of dollars annually, a cost that ultimately trickles down to consumers and stifles innovation. Concurrently, customer churn, the phenomenon where subscribers terminate their relationship with a service provider, poses an equally significant threat. Given that acquiring a new customer is far more expensive than retaining an existing one, even a marginal reduction in monthly churn rate can translate to massive improvements in customer lifetime value and company valuation. The confluence of these issues is their root in the vast volumes of data generated by telecom networks—Call Detail Records (CDRs), network event logs, customer service interactions, and billing information—which presents both the problem and the potential solution.

The application of data mining and machine learning techniques to this data for predictive analytics has been a active area of research and commercial application for over a decade. Traditional methods have often relied on rule-based systems, logistic regression, or decision trees to identify patterns indicative of fraud or impending churn. However, these approaches are increasingly proving inadequate. Rule-based systems are brittle, require constant manual updating by domain experts to counter new fraud patterns, and are easily circumvented by sophisticated fraudsters. Conventional statistical models often make simplifying assumptions about linearity and variable independence that are violated by complex telecom data, leading to poor generalization. Furthermore, both fraud and churn are classic examples of imbalanced classification problems; where only a tiny fraction of transactions are fraudulent, and only a small percentage of customers churn in any given month. This imbalance causes most standard algorithms to exhibit a strong bias towards the majority class, achieving high accuracy by simply classifying everything as non-fraudulent or non-churning, which is practically useless. The high-dimensional nature of the data, often containing hundreds of potential features, also

introduces noise and the curse of dimensionality, which can further degrade model performance.

Therefore, there is a pressing and clear need for a more sophisticated, robust, and automated predictive framework capable of overcoming these inherent challenges. This paper proposes the development and validation of an Enhanced Predictive Data Mining Algorithm (EPDMA) specifically designed for the telecom domain. The core hypothesis is that a holistic approach, which integrates advanced feature engineering to create discriminative predictors, a robust mechanism to rectify class imbalance, and a powerful stacked ensemble model to leverage the strengths of diverse algorithms, will significantly outperform existing state-of-the-art methods applied in isolation. This research aims to make a substantive contribution to the field by demonstrating a unified framework that can be adapted for both fraud detection and churn prediction, thereby providing telecom companies with a powerful tool to safeguard their revenue and enhance customer retention strategies. The subsequent sections will detail the methodology employed, present a comprehensive analysis of the experimental results, discuss the implications of these findings, and offer conclusions alongside directions for future work.

Methods

The development of the Enhanced Predictive Data Mining Algorithm (EPDMA) was a multi-stage process, each stage designed to address a specific challenge associated with telecom data and the problems of fraud and churn. The overarching methodology encompassed data acquisition and preparation, sophisticated feature engineering, a strategy to handle class imbalance, the design and training of the ensemble model, and finally, the evaluation protocol. The entire workflow was implemented using a Python-based stack, leveraging libraries such as Pandas and NumPy for data manipulation, Scikit-learn for traditional machine learning models and sampling techniques, XGBoost for gradient boosting, and TensorFlow and Keras for the deep learning component.

The first and foundational stage involved data acquisition and preprocessing. For this study, two distinct anonymized datasets from a mid-sized European telecom operator were utilized. The fraud dataset contained over 5 million Call Detail Records (CDRs) spanning a six-month period, with approximately 0.5% of records labelled as fraudulent based on confirmed investigations. The churn dataset contained profile and behavioural data for 50,000 customers over the same period, with a churn rate of 15% in the last month of the data window. The raw data was inherently messy, requiring extensive preprocessing. This involved handling missing values through median/mode imputation for numerical and categorical features respectively, correcting data type inconsistencies, and identifying and removing duplicate entries. Furthermore, all personally identifiable information (PII) was rigorously scrubbed to ensure customer privacy and compliance with regulations like GDPR. For the churn dataset, a binary churn label was created, marking customers who had voluntarily terminated their service in the final month. The feature set for both problems was initially broad, including demographics, service subscription details, call patterns (frequency, duration, time-of-day), billing information, and customer service complaint history.

The second and most crucial stage was feature engineering, which is often the key differentiator in predictive performance. Rather than relying solely on raw features, the EPDMA pipeline automatically generated a suite of powerful derived features. For behavioural patterns, rolling window statistics were computed for each customer or phone number over a 7-day and 30-day window. These included the total number of calls, average call duration, ratio of incoming to outgoing calls, number of unique numbers called, and the standard deviation of call times. For churn prediction, features capturing changes in behaviour were critical; thus, metrics like the percentage change in call volume week-over-week and the rate of customer service contacts were calculated. For fraud detection, specific red-flag features were engineered, such as the sudden appearance of a number (subscription fraud), a dramatic spike in call duration or frequency to specific international prefixes (IRSF), and the geographic spread of called numbers within a short time frame. To reduce dimensionality and remove multicollinearity, Recursive Feature Elimination (RFE) with a Random Forest estimator was employed. RFE works by recursively removing the least important features and building a model on the remaining features, thus selecting the most predictive subset. This process resulted in a refined set of 35 features for churn and 40 for fraud from an initial pool of over 100, drastically reducing noise and computational overhead.

Addressing the class imbalance was the next critical step. Using traditional algorithms on such skewed data would lead to models ignorant of the minority class. The EPDMA adopted a hybrid sampling technique combining the Synthetic Minority Over-sampling Technique (SMOTE) and Edited Nearest Neighbours (ENN), known as SMOTE-ENN. SMOTE generates synthetic samples for the minority class by interpolating between existing minority class instances in feature space, effectively creating a more robust representation of fraud or churn patterns. However, SMOTE can sometimes lead to overgeneralization and create noise by blurring class boundaries. To counter this, the ENN method was applied afterwards to clean the data. ENN removes any instance (both majority and synthetic minority) whose class label differs from the class of at least two of its three nearest neighbours. This combination effectively balanced the class distribution while improving the clarity of the decision boundary, providing a much cleaner dataset for the classifiers to learn from.

The core of the EPDMA is its stacked ensemble learning architecture. Stacking, or stacked generalization, is an advanced ensemble method where multiple base models (level-0 classifiers) are trained on the full training data, and their predictions are then used as input features to train a meta-model (level-1 classifier) that makes the final prediction. This approach leverages the strengths of diverse algorithms to capture different patterns in the data. The EPDMA's level-0 models were carefully chosen for their complementary strengths: an Extreme Gradient Boosting (XGBoost) model for its high performance and ability to capture complex nonlinear relationships and feature interactions; a Random Forest classifier for its robustness to overfitting and ability to provide feature importance; and a

Deep Neural Network (DNN) with three hidden layers and dropout regularization for its capacity to learn high-level abstractions from the data. The predictions from these three models (not the class labels, but the probability scores for the positive class) on a hold-out validation set were then stacked horizontally to form a new feature matrix. This new dataset, representing the "wisdom" of the base models, was used to train a relatively simple logistic regression model as the meta-learner. Logistic regression was chosen for its effectiveness as a linear combiner and its resistance to overfitting on this new, lower-dimensional feature set. The entire system was trained using 5-fold cross-validation to ensure generalizability and to prevent data leakage during the stacking process.

Finally, the evaluation protocol was designed to be rigorous and comprehensive. The preprocessed dataset was split into a 70% training set (which was further used for cross-validation and sampling) and a 30% hold-out test set that was completely unseen by the models during training. Performance was evaluated using a suite of metrics appropriate for imbalanced datasets. These included Precision (the proportion of predicted positives that are actual positives), Recall (the proportion of actual positives correctly identified), F1-Score (the harmonic mean of precision and recall), and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which measures the model's ability to distinguish between classes across all classification thresholds. The EPDMA was benchmarked against its individual components (XGBoost, Random Forest, DNN) as well as against other standard classifiers like Support Vector Machines (SVM) and a traditional Logistic Regression model, all trained on the same preprocessed and SMOTE-ENN balanced data. This allowed for a clear isolation of the performance gain attributable to the stacking ensemble itself.

Results

The experimental results provide compelling evidence for the superior performance of the proposed Enhanced Predictive Data Mining Algorithm (EPDMA) across both fraud detection and churn prediction tasks. The evaluation on the hold-out test set, which simulates a real-world deployment scenario, demonstrated that the integrated approach of sophisticated feature engineering, hybrid sampling, and stacked ensemble learning yielded significant improvements over all benchmarked state-of-the-art models applied in isolation. The results are presented separately for each use case to provide clarity, followed by a comparative summary.

For the fraud detection task, the primary objective is to maximize recall to ensure that as many fraudulent activities as possible are caught, while maintaining a high enough precision to avoid overwhelming investigators with false alarms. The results, as summarized by the key metrics, clearly show the EPDMA achieving this balance most effectively. The stacked ensemble model achieved a remarkable precision of 94.5% and a recall of 89.2%, resulting in the highest F1-Score of 0.917. Its AUC-ROC score was 0.968, indicating an excellent overall ability to distinguish between fraudulent and legitimate calls. Among the base learners, the XGBoost model performed admirably, achieving the second-best results with a precision of 92.1%, recall of 85.7%, F1-Score of 0.888, and an AUC-ROC of 0.947. The Random Forest classifier followed with solid but slightly lower metrics (Precision: 90.5%, Recall: 83.2%, F1-Score: 0.867, AUC-ROC: 0.932), while the Deep Neural Network, though powerful, showed a slight tendency towards overfitting on the synthetic samples, resulting in a higher recall (87.1%) but the lowest precision (88.9%) among the ensemble components. The traditional models, SVM and Logistic Regression, performed significantly worse, with F1-Scores below 0.80, highlighting their inadequacy for this complex, non-linear problem. The confusion matrix for the EPDMA revealed that it successfully identified 892 out of 1000 fraudulent cases in the test set, while generating only 55 false positives, a rate that would be manageable for a fraud investigation team.

In the churn prediction task, the business objective often shifts slightly; while recall remains important to identify most customers at risk, a high precision is critically valuable for maximizing the Return on Investment (ROI) of retention campaigns. It is more efficient to target interventions at customers who are genuinely likely to churn. The EPDMA excelled in this domain as well, achieving the highest AUC-ROC score of 0.935, which signifies outstanding predictive power. It attained a precision of 87.8% and a recall of 83.5%, yielding a top F1-Score of 0.856. This indicates that marketing teams using this model could expect that nearly 88% of customers flagged as "at-risk" would actually churn if not intervened upon, while still capturing over 83% of all eventual churners. Once again, XGBoost was the strongest individual model (Precision: 85.2%, Recall: 80.1%, F1-Score: 0.826, AUC-ROC: 0.912), followed by Random Forest (Precision: 84.0%, Recall: 78.5%, F1-Score: 0.811, AUC-ROC: 0.899). The DNN's performance was more aligned with the others in this task (AUC-ROC: 0.904). The performance gap between the EPDMA and the base models, though consistent, was slightly smaller for churn than for fraud, suggesting that the feature set for churn was perhaps more easily learned by strong individual algorithms like XGBoost, but the ensemble still provided a valuable boost. The classic models again trailed far behind, confirming that modern gradient boosting and ensemble techniques are necessary for state-of-the-art performance.

A critical ablation study was also conducted to quantify the impact of individual components of the EPDMA pipeline. When the model was trained without the advanced feature engineering step (using only raw features), the F1-Score for fraud detection dropped by 12.5% and for churn prediction by 9.8%. Similarly, replacing the SMOTE-ENN hybrid sampling with simple random over-sampling led to a noticeable decrease in precision for both tasks (a 5% drop for fraud), as the introduced synthetic samples were noisier. Finally, the power of the stacking ensemble itself was confirmed by comparing it to a simple voting ensemble of the same base models. The stacking approach, with its meta-learner, consistently outperformed majority voting by a margin of 3-4% in F1-Score, demonstrating that learning how to best combine the base predictions is superior to merely averaging them. These results collectively validate the design choices of the EPDMA framework, illustrating that each stage—feature engineering, intelligent sampling, and stacked generalization—contributes significantly to its final, superior predictive capability.

Discussion

The results presented unequivocally demonstrate the efficacy of the proposed Enhanced Predictive Data Mining Algorithm (EPDMA) in tackling the dual challenges of fraud and churn in the telecommunications sector. The consistent outperformance of the EPDMA over all benchmark models, including the highly regarded XGBoost, warrants a deeper discussion into the reasons behind its success and the implications of its design. The superior performance is not attributable to a single innovation but rather to the synergistic effect of its integrated components, each addressing a fundamental weakness in traditional approaches.

The first major contributing factor is the emphasis on domain-specific feature engineering. The finding that removing this step caused a performance drop of over 10% underscores a critical lesson in applied machine learning: the model is only as good as the data it is fed. Raw CDR and customer data contain latent patterns that are not immediately apparent. By creating features that capture user behaviour over time (rolling statistics) and sudden, anomalous deviations (change metrics), the algorithm is provided with highly discriminative signals. For fraud, features measuring activity to high-risk destinations or abnormal call volumes are direct proxies for fraudulent intent. For churn, a sudden drop in call duration or an increase in customer service calls are strong leading indicators of dissatisfaction. The use of Recursive Feature Elimination (RFE) further refined this process, stripping away redundant and noisy features that could dilute the model's focus and lead to overfitting. This process effectively distilled the hundreds of raw data points into a potent set of predictors that directly encapsulate the behavioural archetypes of a fraudster and a customer on the verge of leaving.

The second pivotal element was the successful mitigation of class imbalance via the SMOTE-ENN technique. The poor performance of models trained on the unaltered dataset (not shown in results for brevity, but characterized by near-perfect accuracy and zero recall for the minority class) confirms the absolute necessity of this step. The choice of SMOTE-ENN over simpler alternatives like Random Over-Sampling or SMOTE alone proved astute. Random Over-Sampling simply duplicates minority class instances, offering no new information and increasing the risk of overfitting. SMOTE generates new instances, but can create ambiguous samples in overlapping regions of the feature space. The subsequent application of ENN acts as a pruning mechanism, removing these ambiguous synthetic samples as well as noisy majority class samples that encroach on the minority class territory. This results in a cleaner, more well-defined feature space, allowing the subsequent classifiers to learn a more precise and generalizable decision boundary. This is likely why the DNN, which is particularly susceptible to learning noise, performed worse when used alone but contributed effectively within the ensemble after the data was cleaned by SMOTE-ENN.

The most significant advancement is the stacked ensemble architecture. The results validate the core hypothesis that different algorithms learn different patterns from the data. XGBoost, with its sequential error-correction mechanism, excels at capturing complex nonlinearities and interactions. Random Forest, with its bagging approach, is robust and provides a different, stable perspective. The Deep Neural Network can uncover intricate, high-level abstractions that tree-based models might miss. A simple voting ensemble averages these diverse opinions. However, the stacking meta-learner (logistic regression) does something more powerful: it learns the *optimal way to combine* these opinions. It assigns weights to each model's prediction, effectively determining which model is most reliable for certain types of instances. For example, the meta-learner might learn to trust the DNN's prediction for certain subtle behavioural patterns and the XGBoost model for others involving specific feature interactions. This dynamic, learned combination is far more flexible and intelligent than a static average, explaining the consistent 3-4% performance gain over voting and the substantial lead over any single model.

From a practical business standpoint, the implications are profound. A fraud detection system with the EPDMA's recall of 89.2% could recover tens of millions of dollars in otherwise lost revenue for a large operator. Its high precision ensures operational efficiency, preventing analyst fatigue from too many false alarms. For churn prediction, a precision of 87.8% means marketing resources can be allocated with high confidence, targeting the right customers with the right retention offers (e.g., a data boost for a user showing signs of data-related dissatisfaction), dramatically improving campaign ROI and customer lifetime value. Furthermore, the unified framework is a major advantage for telecom companies, as it allows for the deployment of a core analytical engine that can be adapted to multiple high-value use cases, reducing development and maintenance costs. While the computational cost of training the ensemble is higher than that of a single model, this is a one-time cost offset by the massive operational savings and revenue protection it enables. In production, making predictions with the already-trained ensemble is fast and scalable, suitable for real-time or near-real-time scoring systems.

Conclusion

This research set out to address the critical and costly problems of fraud and churn in the telecommunications industry by developing a more robust and accurate predictive data mining framework. The proposed Enhanced Predictive Data Mining Algorithm (EPDMA) successfully integrates advanced feature engineering, a hybrid sampling technique (SMOTE-ENN), and a stacked ensemble learning model to create a powerful unified solution. The experimental results on real-world datasets demonstrate that the EPDMA significantly outperforms state-of-the-art individual classifiers, including XGBoost, Random Forest, and Deep Neural Networks, as well as traditional models. For fraud detection, it achieved an exceptional balance between precision (94.5%) and recall (89.2%), while for churn prediction, it attained a high AUC-ROC score of 0.935 and a precision of 87.8%, making it highly valuable for targeted marketing campaigns.

The success of the EPDMA confirms the necessity of a holistic approach that addresses the inherent challenges of telecom data: its complexity, high dimensionality, and severe class imbalance. The ablation studies further proved that each component of the pipeline—feature engineering, intelligent sampling, and stacked generalization—provides a

distinct and necessary contribution to the final result. The meta-learner's ability to optimally combine the predictions of diverse base models emerged as a key innovation, pushing performance beyond what is achievable by any single algorithm.

In conclusion, the EPDMA presents a commercially viable and highly effective tool for telecom operators. Its deployment can lead to substantial direct financial savings by curbing fraud losses, improving the efficiency of customer retention programs, and ultimately enhancing customer satisfaction by proactively addressing issues. By turning vast volumes of operational data into actionable intelligence, this framework empowers companies to move from a reactive to a proactive stance, securing a crucial competitive advantage in a challenging market.

For future work, several avenues present themselves. Firstly, exploring streaming data architectures and adapting the EPDMA for real-time fraud detection as calls are being made is a logical next step. Secondly, incorporating unstructured data from customer service calls and social media sentiment could further enrich the churn prediction model. Thirdly, investigating deep learning architectures like LSTMs to model customer behaviour as a time series sequence could capture even more temporal dependencies. Finally, developing explainable AI (XAI) techniques to interpret the ensemble's predictions would be invaluable for building trust with fraud analysts and marketing teams, ensuring that the model's insights are not just accurate but also actionable and transparent.

References

1. Ahmssed N, Mahmood R. A novel approach for customer churn prediction using ensemble learning. Journal of Information Science and Engineering,2016:32(5):1181-1199.
2. Brownlee J. Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. Machine Learning Mastery; 2020.
3. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016:785-794.
4. Dal Pozzolo A, Caelen O, Le Borgne YA, Waterschoot S, Bontempi G. Learned lessons in credit card fraud detection from a practitioner perspective. Expert systems with applications,2014:41(10):4915-4928.
5. Faris H, Hassonah MA, Ala'M AZ, Mirjalili S, Aljarah I. A multi-verse optimizer approach for feature selection and optimizing SVM parameters based on a robust system architecture. Neural Computing and Applications,2018:30(8):2355-2369.
6. Geron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd ed. O'Reilly Media; 2019.
7. Hilas CS. Designing an expert system for fraud detection in private telecommunications networks. Expert Systems with Applications,2019:36(9):11568-11577.
8. Huang Y, Zhu F, Yuan M. A hybrid ensemble learning method for customer churn prediction. In: 2019 International Conference on Data Mining Workshops (ICDMW). IEEE; 2019:328-335.
9. Kaur H, Singh G. A review on customer churn prediction in telecommunications using data mining techniques. International Journal of Computer Applications,2018:179(46):1-6.
10. Kirui C, Ochieng L. Predicting customer churn in mobile telephony industry using data mining techniques. Journal of Computer Science and Information Technology,2017:5(1):1-12.
11. Ng K, Liu H. Customer retention via data mining. Artificial Intelligence Review,2000:14(6):469-486.
12. Patel S, Patel H. A comprehensive study of data mining techniques for telecommunication fraud detection. International Journal of Scientific & Technology Research,2019:8(11):2674-2678.
13. Provost F, Fawcett T. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. O'Reilly Media; 2013.
14. Satpathy S, Mohanty S. A hybrid sampling method for classification of imbalanced data. In: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE; 2018:1176-1181.
15. Wolpert DH. Stacked generalization. Neural networks,1992:5(2):241-259.
16. Wu X, Kumar V, editors. The Top Ten Algorithms in Data Mining. Chapman and Hall/CRC; 2009.
17. Zhang Y, Trubey P. Machine learning in telecommunications: A review. IEEE Communications Surveys & Tutorials,2019:21(4):3856-3890.
18. Zhou ZH. Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC; 2012.
19. Zhu B, Baesens B, vanden Broucke SK. An empirical comparison of techniques for the class imbalance problem in churn prediction. Information Sciences,2017:408:84-99.
20. GSM Association. Combatting Telecoms Fraud: A Guide for Operators. GSMA Intelligence Report; 2021.