International Journal of Research in Engineering

E-ISSN: 2664-8784
P-ISSN: 2664-8776
IJRE 2024; 6(2): 14-20
© 2024 IJRE
www.engineeringpaper.net
Received: 19-06-2024
Accepted: 28-07-2024

Selahadin Nurga Babeta
Department of Computing and Analytics, Information Network Security Administration (INSA), Addis Ababa, Ethiopia

Million Meshesha
Ph.D., School of Information Science, Addis Ababa University, Addis Ababa, Ethiopia

Correspondence
Selahadin Nurga Babeta
Department of Computing and Analytics, Information Network Security Administration (INSA), Addis Ababa, Ethiopia

Telecom airtime credit risk prediction using machine learning

Selahadin Nurga Babeta and Million Meshesha

DOI: https://doi.org/10.33545/26648776.2024.v6.i2a.59

Abstract
This research endeavors to enhance telecom airtime credit risk prediction through the application of machine learning algorithms. For financial stability and customer satisfaction, ethio telecom, the top telecommunications provider in Ethiopia, must effectively manage credit risk. Accurate credit risk prediction can assist the business in identifying clients who are more likely to default on their airtime credit, enabling proactive measures to reduce risks and improve financial performance. The historical customer information included in the dataset for this study includes customer profiles, call records, credit repayment histories, and usage data. Data preprocessing techniques are used before model training to handle missing values, encode categorical variables, and reduce features, ensuring the quality and consistency of the dataset. Machine learning algorithms such as Random Forest, Logistic Regression, Na¨ıve Bayes, and K- Nearest Neighbors (KNN) are leveraged to construct a predictive model under different experimental conditions. After controlling the impact of class imbalance and introducing novel attributes, experimental result shows that Random Forest and Logistic Regression machine learning algorithms exhibit promising results in predicting airtime credit risk. One of the major challenges in this research is dealing with class imbalance in the dataset, where the number of instances of customers who default on their airtime credit is significantly higher than those who do not. To address this challenge, future work should focus on implementing advanced techniques for handling class imbalance, such as synthetic data generation (e.g., SMOTE) and exploring ensemble methods that combine multiple algorithms to improve predictive performance. Additionally, continuously incorporating new and relevant attributes and refining the feature selection process will further enhance the model’s accuracy and reliability.

Keywords: Telecommunication service, machine learning, telecom airtime, credit-risk prediction

Introduction

In today’s highly competitive business environment, the importance of understanding and managing risk has in-creased attention within the scientific community. The concept of” risk” permeates various economic, social, and scientific con- texts, with its prominence notably observed in finance, banking, insurance, and medicine [7]. For businesses, particularly those offering services like telecommunication, identifying and managing risk factors in their processes is crucial for maintaining a competitive edge. The telecommunication sector, characterized by rapid technological advancements and evolving business models, necessitates a comprehensive understanding of risk factors for sustainable success. Managing risks, such as churn analysis, fraud detection, customer segmentation, and optimal use of telecommunications infrastructure, is pivotal in this dynamic market [16]. One specific area of concern within the telecommunications industry is credit risk, which has gained significance in recent years. Credit risk analysis of individual and corporate customers during the activation process has become vital for operational processes. In the context of the telecommunication sector, credit risk translates into potential profit reduction, cash flow shortfalls, and financial challenges that could lead to bankruptcy [7, 16]. The provision of airtime credit, a service allowing users to obtain short-term airtime loans, has emerged as a strategic offering for telecommunications companies. Customers in need of immediate airtime but unable to purchase recharge cards can avail this service, contributing to increased customer satisfaction and average income per user [6]. In Ethiopia, ethio telecom introduced an airtime credit service which has witnessed significant utilization, with millions of users accessing it monthly and contributing substantially to organizational revenue [2]. However, managing credit risk associated with airtime loans poses challenges, particularly in identifying eligible customers and predicting their creditworthiness. To ad- dress this, machine learning, a subset of artificial intelligence, has gained prominence.

Machine learning techniques offer the capability to analyze vast datasets and identify patterns that may elude human observation. In the telecommunications sector, machine learning has been employed for tasks such as enhancing marketing campaigns, fraud detection, and network optimization [13]. This study focuses on utilizing machine learning algorithms to predict airtime credit risk in the telecommunications industry, with a specific emphasis on ethio telecom. The motivation for this research stems from the industry's rapid evolution, the increasing reliance on data- driven decision-making, and the imperative to manage credit risk effectively. As 5G and Internet of Things (IoT) applications generate substantial data, the development of predictive models becomes critical for informed risk assessment and decision-making [9]. The problem attempted in this study revolves around the effective prediction of airtime credit risk, which requires the identification of suitable attributes and the selection of appropriate machine learning algorithms. While previous studies have attempted to predict airtime credit risk using supervised machine learning algorithms, such as those by Oliyad [17] and Shashu [2], these efforts were hindered by the lack of consideration of potentially relevant customer usage data attributes and advancements in service technology within the telecommunications industry. This study seeks to address these limitations by identifying suitable attributes, selecting appropriate machine learning algorithms, and evaluating the performance of optimal models in predicting airtime risk. Specifically, the research aims to answer the following questions: What attributes and machine learning algorithms are suitable for airtime credit risk prediction? By answering this question, the study aims to contribute to a better understanding of how airtime credit services can be offered effectively to customers, ultimately enhancing credit risk management practices in the telecommunications sector. The findings of this research hold significant implications for telecommunications companies, such as ethio telecom, in managing credit risk and making informed decisions regarding airtime credit services.

Related works

In the realm of credit risk assessment, the literature provides valuable insights into the application of machine learning techniques, especially in the context of the telecommunications sector. Foreign scholars have made significant contributions to understanding and managing credit risk in this industry, utilizing diverse methodologies and models. Monika and An- drzej [16] delved into the application of data mining methods in the telecommunication sector for credit risk prediction and management. Their study emphasized the importance of credit risk management to prevent bad debt and financial losses. Decision trees and variable importance measures are employed to construct models for individual and business customers, based on activation data and payment behavior. The authors recommend further exploration of activation models and different split criteria for predicting customer churn. Bernard *et al* [6] explored the complexities of airtime credit services, particularly the mechanisms where Mobile Network Operators (MNOs) take on the risk of defaulted loans. The research employed binary classifiers, including decision trees, logistic regression, and random forests, to predict loan out- comes. The study revealed the challenges posed by limited data exchange between MNOs and lenders, emphasizing the need for privacy protection. While Random Forest emerges as the top classifier, the study underscores the necessity of considering demo-graphic information for a comprehensive examination. Dengov [5] focused on credit risk analysis for Russian telecommunications companies, employing statistical techniques such as logistic regression analysis. The study high- lighted the significance of accurately assessing credit risk in a rapidly expanding telecommunications sector. Den-gov's model, based on financial ratios, underscored the importance of data quality and preprocessing for precise statistical models. However, the study acknowledged limitations in the exclusive reliance on financial ratios and suggests the inclusion of other relevant factors for a comprehensive credit risk analysis. Oliyad [17] investigated air-time credit risk, recognizing it as a means to enhance customer satisfaction but acknowledging the challenges in debt repayment. The study applied data mining techniques, such as Naive Bayes, Multilayer Perceptron, Logistic Regression, and J48 Decision Tree. The J48 Decision Tree outperforms other classifiers, achieving an accuracy rate of 98.56 percent. The study identified data usage as the main attribute with predictive power and suggests the incorporation of subscriber loan history for further model improvement. Shashu [2] focused on machine learning-based mobile airtime credit risk prediction in the context of value-added services (VAS) provided by Mobile Network Operators (MNOs). The study utilizes four machine learning algorithms—decision tree, logistic regression, random forest, and multilayer perceptron. J48 decision tree performs best among them. While the study enhances prediction accuracy with a combined feature set, it lacks detailed information on feature identification and data collection methodologies. These works collectively contribute to the study by addressing the challenges and opportunities in credit risk prediction within the telecommunications sector. They underscore the importance of refining models, considering diverse factors such as They underscore the importance of refining models, considering diverse factors such as customer behavior, transaction history, and demographic information, and addressing the evolving dynamics of the industry for effective credit risk management.

Methodology

Methodological Framework

This research undertakes a rigorous methodological approach to develop and evaluate machine learning models for predicting airtime credit risk in the Ethiopian telecommunications sector. As python emerges as the preferred tool, offering a versatile ecosystem for machine learning, its capabilities are harnessed for data manipulation, preprocessing, constructing a model, and evaluation. In this study, an experimental re- search design is adopted, providing a systematic approach to assess the performance of different machine learning models. Drawing inspiration from established methodologies [14, 10] the study progresses through various stages, such as problem identification, data collection, preprocessing, and model training, each of them enables to ensure a comprehensive and methodical exploration of credit risk prediction. The details of each step is shown below in figure 1. As noted in the study each of them is crucial for the successful execution of the research work.

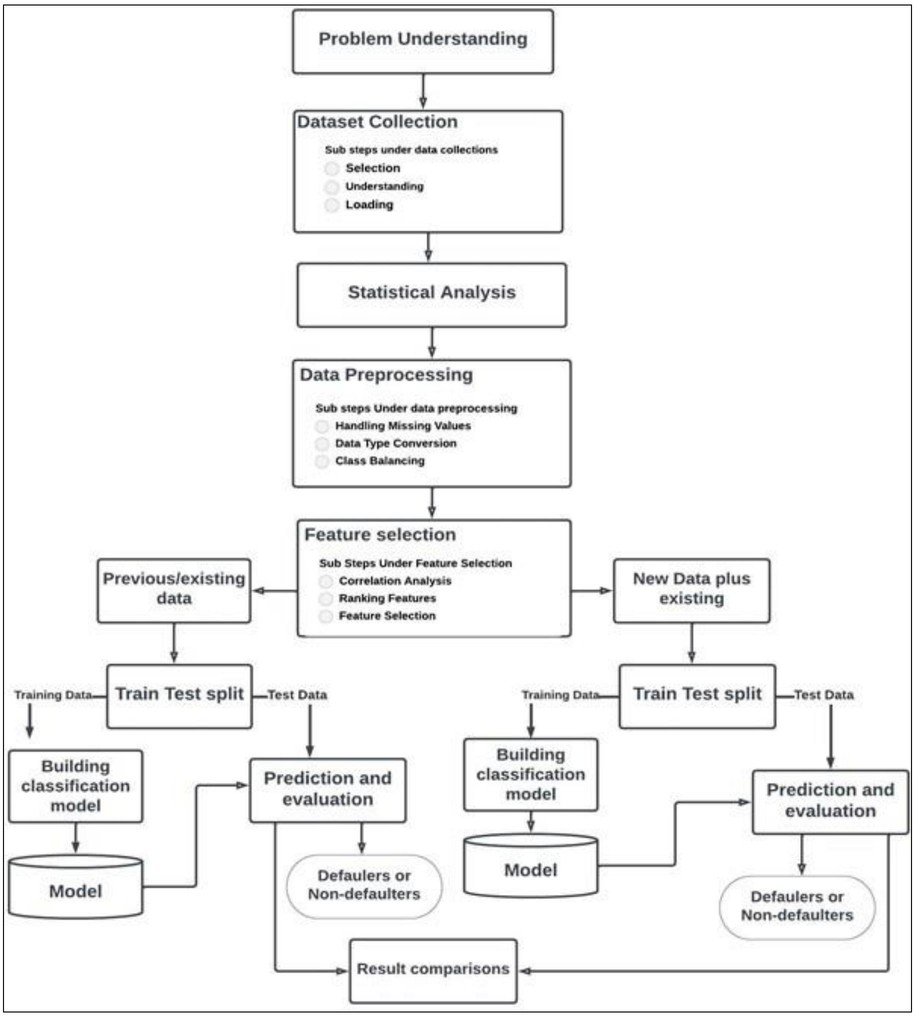

Fig 1: The proposed architecture

Problem Understanding

This research delves into understanding the complexities of airtime credit risk in the telecommunications industry, aiming to identify key challenges and solutions. Airtime credit presents significant risks, including customer default and revenue loss for mobile network operators like ethio telecom. To mitigate these risks, effective tools for predicting credit risk are essential [5, 8, 3]. Supervised machine learning algorithms offer promise in this regard but remain largely un- explored in the context of airtime credit. Telecommunications data, including call logs and customer usage patterns, can be leveraged for credit scoring and assessing creditworthiness. Recharge datasets provide insights into customer behavior and usage patterns, aiding in risk prediction [1, 4, 11, 15]. This study investigates customer behavior, transaction history, and other variables to develop effective predictive models for anticipating credit risk in airtime credit users. By understanding these factors, mobile network operators can better manage credit risk exposure and protect revenue streams.

Data Collection Method

The dataset used for this study was obtained from the company's business administration systems, specifically focusing on subscriber data spanning a three-month period from October to December 2022. This limited timeframe was dictated by ethio telecom's data retention policy, which mandates information retention for a maximum of 45 days to three months before archival.

Customer Profile Data: The first dataset, derived from one of the telecom database systems, comprises customer profile data. This dataset encompasses details maintained about telecom customers, including customer type, education, gender, age, religion, and service number. While this data offers insights into customer behavior and preferences, ethical considerations and privacy regulations necessitated the exclusion of the" service number" attribute from the analysis, despite its role as a primary key for dataset integration.

Loan Information Dataset: The second dataset obtained from the telecom database system pertains to loan information. This dataset encompasses data generated by the telecom operator regarding loans or credit facilities provided to customers, including loan amounts, repayment periods, due dates, outstanding balances, and payment histories. It serves as a crucial tool for assessing customer creditworthiness and managing credit risk, enabling operators to identify customers at higher risk of default and implement appropriate risk mitigation strategies.

Usage Detail Dataset: Another dataset collected from the telecom database system is the customer usage detail dataset. This dataset records customer usage of telecommunication services, including call duration, data usage, and text message volume. It provides valuable insights into customer behavior, allowing operators to identify trends, optimize net- work capacity, develop new services, and tailor marketing efforts to better meet customer needs.

Call Detail Dataset (Recharge Dataset): The fourth dataset obtained from the telecom comprises call detail data, specifically focusing on mobile phone recharge transactions. This dataset includes details such as the phone number being recharged, recharge amount, and transaction date and time. It offers insights into customer recharge behavior and us- age patterns, informing operational decisions and marketing strategies. Overall, the combination of these datasets provides a comprehensive understanding of customer behavior, usage patterns, and credit risk factors, enabling ethio telecom to make informed decisions and optimize its services to meet customer needs effectively.

Data Construction

Merging data is essential to combine information from various sources for comprehensive analysis. However, dealing with heterogeneous and large-scale raw data presents challenges. The researcher merged ethio telecom's datasets using Panda's data analysis library in python programming language. The datasets varied in size and shape, with the customer usage data being too large at 36GB. To handle this, the researcher split it into three parts, removed duplicates, and concatenated them into one data frame. Similarly, the loan information dataset, with 95,786,870 rows, was challenging to process despite using a high-capacity server. After deduplication, it was reduced to 30,980,644 rows. The recharge information dataset, with 5,259,892 rows, was merged with the loan information based on service numbers to create a dataset with 131,698,319 rows. Additionally, the customer profile data underwent deduplication, resulting in 6,174,513 rows. Using service numbers as the primary key, all datasets were linked to create a unified table with a random sample of 2,000,000 instances, serving as the basis for further analysis.

Statistical Analysis for Data Reduction

A detailed statistical analysis is conducted to understand the data's characteristics. Data types, correlations between attributes, and potential redundancies are explored to inform subsequent preprocessing steps effectively. In this study, the pairwise Pearson correlation coefficient of attributes are examined to identify and remove highly correlated attributes, mitigating multicollinearity [12]. The correlation analysis revealed strong correlations between most attributes. At- tributes" Initial loan amount" and" Initial loan poundage" exhibited a correlation value of 1, indicating identical values in each instance. Similarly," INTER-USG-MINUTE" and" INTER-AMT-ETB" showed a strong correlation of 0.97. Conversely, attributes such as" LOAN-BALANCE-TYPE"," LOAN-PENALTY"," LOAN-PENALTY-LEFT", and" SMS-
INTER-FEE-ETB" displayed no correlation, as their values remained consistent across all instances. Also, a careful consideration is given to data reduction, focusing on relevant instances. The data reduction method refines the dataset by deleting irrelevant or redundant data in order to make it more suitable for analysis or modeling. Initially, filtering criteria are used to choose keep instances and features that correspond to the study objectives. For example, in the context of customer data, this might indicate omitting 'Enterprise' customers or focusing solely on those with active service subscriptions. Furthermore, constant columns or features with equal values across all records are discovered and removed because they provide no discriminatory power or useful insights. This filtering procedure produces a more compact and concentrated dataset, which improves the rate and effectiveness of following analytical tasks or machine learning algorithms.

Data Pre-processing

Efficient data preparation is crucial for successful machine learning outcomes [1]. The research incorporates steps such as filtering, handling missing values, encoding categorical features, and balancing the dataset using oversampling and un- der sampling techniques. An assessment of missing values in relevant features is conducted. After reviewing the dataset, it is discovered that some attributes had missing values. Attributes such as 'CUSTOMER-AGE', 'GENDER' and 'EDUCATION' had 3877, 3727 and 126276 missing values, respectively. CUSTOMER-AGE is an integer data type, but 'GENDER- NAME' and 'EDUCATION-NAME' are nominal data types. To handle the missing values, a forward fill method was used. Forward filling is the process of inputting missing data using the previous observed value.

Encoding Categorical Features

Categorical features are converted into numerical values through encoding, facilitating seamless integration into machine learning algorithms. This step is pivotal for ensuring the algorithms can effectively process the data. After all of the categorical attributes were converted to numerical values, the researcher created a target variable called" Loan-Balance" by calculating the difference between the initial loan amount (INIT-LOAN-AMT) and the repayment amount (REPAY-AMT). In the resulting dataset, instances with a Loan- Balance value of zero were labeled as non-defaulters, and those with a value of non-zero were labeled as defaulters. After labeling the target variable" Loan balance", it was found that 73 percent of cases categorized as 1 (defaulters) belonged to the majority class, while 23 percent of instances labeled as 0 (non-defaulters) belonged to the minority class. Acknowledging this unbalanced distribution, the researcher addressed it and reduced its possible effect on model performance by applying data balancing approaches consistently throughout all experiments.

Forming data sets

Matrices of features and the target variable (Loan-Balance) are created, setting the stage for model building. This involves mapping the difference between the initial loan amount and the repayment amount, aiding in class labeling. The dataset is also split into training and testing sets, with 80 percent dedicated to training and 20 percent for testing. This partitioning ensures the model is trained on a substantial portion of the data while retaining a separate set for unbiased evaluation. Towards training and model creation, various classification algorithms, are considered.

Experimental Results

This study employed various machine learning supervised algorithms, such as Random Forest, Logistic Regression, Naive Bayes and K-Nearest Neighbor (KNN) to analyze a dataset related to default prediction. The dataset underwent different preprocessing techniques, including class balancing, the introduction of novel attributes, and attribute removal. The performance of each algorithm was evaluated across these experiments using recall, precision, accuracy, and F- measure. Confusion matrix plays a pivotal role in quantifying the model’s effectiveness.

Experimenting the effect of data balancing

To see the effect of data balancing, an experiment is conducted using under-sampling and over-sampling. Table 1 below summarizes the experimental result for each algorithm.

The Random Forest algorithm exhibited exceptional accuracy, reaching 99.9 percent without class balancing. Under-sampling and over-sampling techniques maintained high accuracy with accuracy of 99.1 percent and 99.9 percent, respectively. The second-best result is obtained by logistic regression, with unbalanced dataset, achieved 93.9 percent accuracy. After under-sampling, accuracy increased to 98.2 percent, and over-sampling led to 97.97 percent accuracy. Both methods significantly improved model performance. On the other hand, Na¨ıve Bayes showed the least performance of 82.67 percent accuracy without class balancing. Under- sampling and over-sampling led to decreased accuracy to 68.23 percent and 68.1 percent, respectively.

Table 1: Summary of the Experimental Result Before and After Sampling Data Set

No	Algorithm	Experimental Setup	Accuracy
1	Random Forest	Before Applying Class Balancing	0.999
2	Random Forest	Balanced Target Data (Under-sampling)	0.991
3	Random Forest	Balanced Target Data (Over-sampling)	0.999
4	Logistic Regression	Before Applying Class Balancing	0.939
5	Logistic Regression	Balanced Target Data (Under-sampling)	0.982
6	Logistic Regression	Balanced Target Data (Over-sampling)	0.979
7	Na¨ıve Bayes	Before Applying Class Balancing	0.826
8	Na¨ıve Bayes	Balanced Target Data (Under-sampling)	0.682
9	Na¨ıve Bayes	Balanced Target Data (Over-sampling)	0.680
10	KNN	Before Applying Class Balancing	0.903
11	KNN	Balanced Target Data (Under-sampling)	0.731
12	KNN	Balanced Target Data (Over-sampling)	0.898

Experimenting the effect of constructing new Attributes from Existing Attributes

An experiment is also conducted to investigate the effect of constructing new attributes from existing attributes on the performance of classification algorithms. Table 2 below compares predictive performance using newly constructed attributes versus existing attributes.

Table 2: Performance Comparison of Classification Algorithms on Different Datasets

Dataset Used	Algorithm	Accuracy	Class	Precision	Recall	F1-Score
Dataset with new attributes	Random Forest	98.40%	0	0.95	0.99	0.97
			1	1.00	0.98	0.99
	Logistic Regression	94.70%	0	0.91	0.90	0.90
			1	0.96	0.96	0.96
	Na¨ıve Bayes	68.4%	0	0.47	1.00	0.64
			1	1.00	0.57	0.72
	KNN	97.3%	0	0.93	0.97	0.95
			1	0.99	0.97	0.98
Dataset with existing attributes	Random Forest	69.2%	0	0.41	0.27	0.32
			1	0.75	0.85	0.80
	Logistic Regression	72.5%	0	0.25	0.00	0.00
			1	0.73	1.00	0.84
	Na¨ıve Bayes	72%	0	0.29	0.01	0.03
			1	0.73	0.99	0.84
	KNN	70.3%	0	0.45	0.35	0.40
			1	0.77	0.84	0.80

The introduction of new attributes improved the accuracy of each classification algorithms. Random Forest achieved 98.4 percent accuracy as compared to using only existing attributes which register 69.2 percent accuracy. Logistic Regression, Na¨ıve Bayes, and KNN also showed varying degrees of improvement with the novel new attributes.

The purpose of this experiment is to evaluate how classification algorithms performed when specific but important attributes that were used to create class labels were excluded. The study aimed to determine the significance of these features in precisely forecasting credit risk by deleting them and com- paring the outcomes with the experiment carried out prior to their removal. This investigation offered insightful information about the significance of particular characteristics in improving classification algorithms’ performance and their function in accurately predicting credit risk. After removing attributes used in creating class labels, Random Forest accuracy decreased to 80.4 percent, emphasizing the importance of these attributes. Logistic Regression and Na¨ıve Bayes showed reduced performance, while KNN maintained reasonable accuracy (75.3 percent). The following table presents the performance of algorithms before and after removal of attribute that were used for class labeling purposes:

Table 3: Performance Comparison of Classification Algorithms Before and After Attribute Removal

Dataset Used	Algorithm	Accuracy	Class	Precision	Recall	F1-Score
Before Removing the attributes	Random Forest	99.9%	0	0.99	1.00	0.99
			1	1.00	0.99	0.99
	Logistic Regression	93.9%	0	0.88	0.91	0.89
			1	0.96	0.95	0.96
	Na¨ıve Bayes	82.6%	0	0.77	0.53	0.63
			1	0.84	0.94	0.89
	KNN	90.3%	0	0.82	0.82	0.82
			1	0.93	0.93	0.93
After Removing the attributes	Random Forest	80.4%	0	0.65	0.63	0.64
			1	0.86	0.87	0.87
	Logistic Regression	76.4%	0	0.65	0.32	0.43
			1	0.78	0.93	0.85
	Na¨ıve Bayes	68.7%	0	0.47	0.91	0.62
			1	0.95	0.60	0.74
	KNN	75.3%	0	0.55	0.52	0.54
			1	0.82	0.84	0.83

Overall, experimental results showed that, random forest consistently outperformed other algorithms, especially when novel attributes were introduced. Logistic Regression and Na¨ıve Bayes were sensitive to class distribution changes, while KNN demonstrated resilience. Hence, I suggest the use random forest predictive model for predicting airtime credit in telecommunication service provision.

Implications and Considerations

The study underscores the importance of tailored approaches in handling class imbalances. The choice of class balance technique and the inclusion of novel attributes significantly impact the model’s ability to predict credit risk accurately. While Random Forest excelled in its original form, other algorithms benefited from specific class balance techniques. This highlights the need for a nuanced selection process, considering algorithmic idiosyncrasies. The inclusion of novel derived attributes proved beneficial across algorithms, suggesting avenues for further exploration and feature engineering. This study makes a substantial contribution by identifying critical characteristics that indicate airtime credit risk. The models that are developed have a high level of prediction accuracy since we use a complete data from client profiles, use patterns, loan histories, and recharge behavior. Unlike earlier research, the distinguishing feature of these variables is their remarkable predictive power for determining airtime credit risk. Furthermore, this work contributes to the field of credit risk prediction by focusing on airtime credit services in the telecom industry, providing insights that differ from typical financial lending organizations. The study improves predictive performance by assessing various machine learning algorithms, such as Random Forest, Logistic Regression, Na¨ıve Bayes, and K-Nearest Neighbors. In addition, the study systematically investigates different class balancing methodologies, such as under- and over-sampling, to improve predictive capacities, particularly in the realm of airtime credit risk prediction.

Conclusion

This paper presents the efficacy of machine learning algorithms for airtime credit risk prediction. The experimental results underscore the effectiveness of Random Forest and KNN in achieving high accuracy, showcasing their potential for robust risk assessment in airtime credit services. Notably, the study reveals the critical role of class balancing techniques, with over-sampling demonstrating significant improvements in model performance. The introduction of novel attributes enhances predictive accuracy, emphasizing the importance of feature engineering in credit risk modeling. The comparative analysis of algorithms using new versus existing attributes provides valuable insights into the impact of attribute selection on predictive performance. Moreover, this research contributes to the broader discourse on credit risk prediction by providing a nuanced understanding of machine learning models in the specific context of airtime credit services. As machine learning continues to evolve, the findings of this study offer practical implications for credit risk management in the telecommunications sector. The insights gained can guide the refinement of credit scoring systems, ultimately fostering more informed lending decisions and reducing the risks associated with airtime credit services. The research lays the foundation for further investigations and advancements in the intersection of machine learning and credit risk assessment, propelling the field towards more robust and accurate predictive models.

Recommendations

This research results provide domain experts in the telecom industry with valuable insights and tools to improve their airtime credit risk management. The predictive models, attribute significance, and performance metrics obtained from our study enable domain experts to assess risks, make informed decisions, develop early warning systems, segment customers, and optimize collection strategies effectively. By leveraging this research findings, domain experts can enhance their credit risk management practices, minimize defaults, and improve financial performance The following suggestions are proposed based on the airtime credit risk prediction using machine learning conducted in this study.

Implement machine learning models: This method performed remarkably well in predicting the risk of airtime credit. Therefore, it is advised to add a machine learning model as a credit risk management system to be used by ethio telecom. Ethio telecom can improve its risk assessment capabilities and make better decisions about credit extension by utilizing the strengths of machine learning models, such as its capacity to handle complex data and provide accurate predictions.

- Update and improve the predictive model frequently: As consumer behavior and market dynamics change over time, it is crucial to routinely update and improve the predictive model. This can be accomplished by adding fresh data sources, identifying new patterns, and optimizing the machine learning algorithms. ethio telecom can guarantee the effectiveness and relevance of its airtime credit risk prediction system by staying up to date with the most recent information and continuously improving the model.
- Monitor model performance and carry out periodic evaluations: It's crucial to regularly evaluate the predictive model's performance. This involves evaluating the model's accuracy, precision, recall, and F1 scores as well as contrasting the predictions with actual instances of credit default. In order to maintain the model's effectiveness and reliability in predicting airtime credit risk, regular evaluations help identify any potential flaws or areas for improvement.

Future works

Exploration of advanced machine learning methods: Although this study used well-known machine learning algorithms like Random Forest, Logistic Regression, Naive Bayes, and K-Nearest Neighbors, there are many advanced algorithms available that can uncover new insight that were not captured by the algorithm used. Future studies might examine how well other algorithms, like gradient boosting, support vector machines, or deep learning models like neural networks, perform. Comparing how well these algorithms predict airtime credit risk can reveal new information and possibly result in models that are more reliable and accurate.

Conflict of Interest Statement

The author declares that there are no conflicts of interest regarding the publication of this paper. The research was conducted independently, with guidance from an advisor, Million meshesha (PHD), and evaluation by Martha Yifru (PHD) and Tibebe Beshah (PHD). None of the individuals involved have any financial or personal relationships that could inappropriately influence or bias the research and its findings.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Million Meshesha, for his invaluable guidance, support, and encouragement throughout this research. His expertise and insights were instrumental in shaping the direction of this study, and his feedback greatly contributed to the quality and depth of this work. I am also sincerely grateful to Dr. Martha Yifru and Dr. Tibebe Beshah for their thoughtful evaluations and constructive feedback, which significantly enhanced the final manuscript. Their perspectives and recommendations were critical in refining the methodology and analysis presented in this research. I would like to extend my thanks to ethio telecom for providing access to the data necessary for this study. This research would not have been possible without their cooperation and support. Additionally, I would like to acknowledge and thank Seble Hailu (PhD) from INSA for her invaluable support in the process of publishing this paper.

Conflict of interest: Not applicable.

Funding Information: Not applicable.

Author Contribution: The sole author, Selahadin Nurga, conducted the research, analyzed the data, and wrote the manuscript. The advisor, Million Meshesha (PHD), provided guidance and supervision throughout the research process.

Data Availability Statement: The datasets used and/or analyzed during the current study are confidential and are not publicly available due to privacy concerns. However, they are available from the corresponding author on reasonable request, subject to restrictions.

Research Involving Human and/or Animals: Not applicable.

Informed Consent: Not applicable.

References

1. Astorino A, Gorgone E, Gaudioso M, Pallaschke D. Data preprocessing in semi-supervised SVM classification. Optimization. 2011;60(1-2):143-151.
2. Berhe SB. Machine learning-based mobile airtime credit risk prediction using customer profile and loan information. Telecommunication Engineering. 2021.
3. Björkegren D, Grissen D. Behavior revealed in mobile phone usage predicts loan repayment. arXiv preprint arXiv:1712.05840. 2017.
4. Björkegren D, Grissen D. Behavior revealed in mobile phone usage predicts loan repayment. arXiv preprint arXiv:1712.05840. 2017.
5. Dengov V, Tulyakova I. Credit risk analysis for the telecommunication companies of Russia: Statistical model. In: 2nd International Multidisciplinary Scientific Conference on Social Sciences and Arts (SGEM 2015). Albena: Bulgarian Academy of Sciences; 2015.
6. Dushimimana B, Wambui Y, Lubega T, McSharry PE. Use of machine learning techniques to create a credit score model for airtime loans. J Risk Financial Manag. 2020;13(8):180.
7. Lyra M, Onwunta A, Winker P. Threshold accepting for credit risk assessment and validation. J Bank Regul. 2015;16:130-145.
8. Ma L, Zhao X, Zhou Z, Liu Y. A new aspect on P2P online lending default prediction using meta-level phone usage data in China. Decis Support Syst. 2018;111:60-71.
9. Mahmoud H, Ismail T. A review of machine learning use-cases in the telecommunication industry in the 5G era. In: 2020 16th International Computer Engineering Conference (ICENCO). IEEE; 2020. p. 159-163.
10. Nasteski V. An overview of the supervised machine learning methods. Horizons. 2017;4(51-62):56.
11. Ots H, Liiv I, Tur D. Mobile phone usage data for credit scoring. In: Databases and Information Systems: 14th International Baltic Conference, DB&IS 2020, Tallinn, Estonia, June 16-19, 2020, Proceedings 14. Springer; 2020. p. 82-95.
12. Saarela M, Jauhiainen S. Comparison of feature importance measures as explanations for classification models. SN Appl Sci. 2021;3(2):272.
13. Sarker IH. Machine learning: Algorithms, real-world applications and research directions. SN Comput Sci. 2021;2(3):160.
14. Sen PC, Hajra M, Ghosh M. Supervised classification algorithms in machine learning: A survey and review. In: Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018. Springer; 2020. p. 99-111.
15. Sogunro AB, Obiwuru TC, Olaniyan SM, Olaiya IK, Oluwole OA, Ayorinde RO. Trend and pattern of advanced airtime/data lending and its probability distribution on Nigerian telecommunication network. UNILAG J Bus. 2020;6(1):96-113.
16. Szczerba M, Ciemski A. Credit risk handling in the telecommunication sector. In: Industrial Conference on Data Mining. Springer; 2009. p. 117-130.
17. Tarekegn O. Application of data mining technique for predicting airtime credit risk: The case of Ethio Telecom. 2019.