Revitalizing telecoms customer loyalty in advanced analysis of churn prediction with machine learning: A case study of Econet

Kudzai Marutsi, Monica Gondo
Department of Information Sciences and Technology, Harare Institute of Technology, Harare, Zimbabwe

***Abstract*—The researcher is objective to develop an appropriate model which helps in revitalizing telecoms industry customer loyalty in advanced analysis of churn prediction with machine learning. The researcher notes the high rate of churn within the telecoms industry in Zimbabwe due to factors such as pricing, service quality, network connectivity, customer support effectiveness and competitive offerings by other service providers in the market. Data was collected from the telecoms industry to find the major causes of churn prediction and devise appropriate practices to reduce churn and retain and attract customers. The researcher engaged relevant people from relevant departments with the main thrust of obtaining the needed research data.**

To effectively reduce churn within the telecoms industry the researcher concluded that, it is imperative for service providers to enhance customer service quality, offer competitive pricing and promotions, improve network reliability and coverage. Innovation and introduction of new services regularly is crucial to maintain customer loyalty. It is also crucial to use predictive analytics to identify and retain at-risk customers.

Three models; Logistic Regression, Decision Tree and Random Forest were tested by the researcher to determine the effective model that the telecoms industry in Zimbabwe can use to reduce customer churn. Among the tested models, the Random Forest model demonstrated highest accuracy and was recommended to reduce churn rate in the telecoms industry.

I. INTRODUCTION

In the telecoms industry, maintaining customer loyalty is paramount for sustained growth and profitability. One of the significant challenges faced by telecom companies is the high rate of customer churn, where customers switch to competitors or discontinue services altogether. This presents corporate challenges that hinders profitability, growth and attainment of strategic objectives. Traditional methods of churn prediction often fall short in accurately identifying at-risk customers in a timely manner, leading to revenue loss and diminished customer satisfaction. With recent advances in data analytics, many forms of Customer Relationship Management (CRM) systems have been embedded as data analytical methods and these have become the focus of many studies and practices. Such analytical methods pay much attention to customer-centric approaches over product-centric approaches.

As the churn rate, continue to increase within the telecoms industry in Zimbabwe, the major question is, what are the factors contributing most significantly to customer churn within the telecoms industry? With the evolution of technology, a question can be posed; what are the machine learning models that can accurately predict customer churn based on historical data, customer demographics, usage patterns, and service interactions and how best can we implement a scalable framework for real-time monitoring of churn risk, allowing telecom companies to take proactive measures to retain customers before churn occurs?

II. DATA COLLECTION AND PRE-PROCESSING

A. Data Sources

The dataset was sourced from a telecommunications company and contains customer demographics, service usage pat- terns, billing information, and interaction history. The dataset includes 7,043 customer records and 21 attributes, which serve as features for model training.

B. Data Cleaning

Data cleaning involved several steps:

- Handling Missing Values: Missing values were imputed using appropriate techniques based on the data type (mean for continuous variables, mode for categorical variables).
- Outlier Detection and Treatment: Outliers were identified using z-scores and either capped or removed.
- Consistency Checks: Ensured consistency across different datasets by standardizing formats and resolving discrepancies

C. Feature Engineering

Feature engineering was performed to enhance model performance. Key features created included:

- Tenure: Duration a customer has been with the company.
- Monthly Charges: Average monthly expenditure.
- Total Charges: Total expenditure over the customer's tenure.
- Service Usage: Patterns in the usage of services like voice, SMS, and data.
- Customer Support Interaction: Frequency and type of customer support interactions.

III. METHODOLOGY

A. Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution and relationships of variables. Visualization techniques like pie charts, donught charts, graphs and correlation matrices were used to uncover patterns and trends relevant to churn.

Research Methods

Data collection was critical to draw relevant meaningful conclusions to this research project. To have a solid base for the complete result, an exploratory method of design was used. While using the explorative research design, a clear picture on the causes of churn within the telecoms industry in Zimbabwe was uncovered. Through explorative research design, the researcher gathered relevant data on what caused churn and the challenges of customer loyalty in the telecoms industry. In order to carry out an effective explorative research, the researcher used qualitative research approach.

Sample and Sample Size

According to Webster cited in Mugo (2023), a sample is a finite part of a statistical population whose properties are studied to gain information about the whole population. The most critical issue in sampling is representatives. Samples were used by the researcher because they save time and resources. While selecting the sample, the research considered factors such as the number of the targeted population, methods to be used to analyze gathered data, the hypothesis of the study and the instruments to be used to collect data.

Sample Size

Sample size references the total number of respondents included in a study, and the number is often broken down into sub-groups by demographics such as age, gender, and location so that the total sample represents the entire population (Kibuacha, 2021). Saunders et al (2019) points out that, sample size is determined based on a 95% confidence rate interval, an estimate of margin of error and the total population which the sample was to be drawn.

According to Kibuacha (2021) at least 33% of the population under study be used as sample size. Therefore, the researcher selected 4 Technology Services Managers, 2 Risk Officers, 2 Enterprise Data Engineers, and 7 Customer Service Managers. The sample size consists of 15 employees from Econet which consist of 43% of the total population.

Sampling Techniques

Sampling can be classified into probability and non-probability sampling (Panneerselvam, 2023). For the purpose of this research, the researcher used probability sampling. According to Murphy (2019), the approach does not apply to all studies because there are certain conditions to be met. Probability sampling gives an equal opportunity of selected to each department. The researcher randomly selected departments to sample: Technology Services, Enterprise Data Engineers, Customer Services and Risk Management.

B. Model Selection

We explored various machine learning algorithms:

1. Logistic Regression: A Baseline Model For Binary Classification
2. Decision Trees: For Their Interpretability

3. Random Forest: To Handle Overfitting Issues In Decision Trees
4. Gradient Boosting Machines (Gbm: For Their Ability To Improve Accuracy
5. Support Vector Machines (Svm: To Find Optimal Separating Hyperplanes
6. Neural Networks: For Capturing Complex Patterns In The Data

A. Model Training

Decision Tree Construction:

Each decision tree *Ti* in the Random Forest is built using a subset of the training data and a random subset of features. Here's algorithm used for splitting the criterion

$$\Delta I(m, \mathcal{F}_{\text{sub}}, \theta) = I(m) - \left(\frac{N_{\text{left}}}{N_m} I(m_{\text{left}}) + \frac{N_{\text{right}}}{N_m} I(m_{\text{right}}) \right)$$

where θ is the threshold, N_m is the number of samples at node m, N_{left} and N_{right} are the number of samples in the left and right child nodes, and $I(\cdot)$ is the impurity measure.

Ensemble of Trees (Random Forest):

Bootstrap Sampling: Random Forest uses bootstrap sampling to create multiple datasets *Di* (each of size *N*) from the original training data *D*. Each dataset *Di* is used to train a separate decision tree *Ti*.

- **Aggregate Prediction**: For prediction:
  - For classification: Each tree T_i predicts the class label y_i for a new instance $\mathbf{x}$. The final prediction $\hat{y}(\mathbf{x})$ is typically determined by majority voting:

$$\hat{y}(\mathbf{x}) = \text{mode}\{y_i \mid T_i(\mathbf{x})\}$$

  - For regression: Each tree T_i predicts a continuous value y_i. The final prediction $\hat{y}(\mathbf{x})$ is usually the average of all predictions:

$$\hat{y}(\mathbf{x}) = \frac{1}{B} \sum_{i=1}^{B} T_i(\mathbf{x})$$

where B is the number of trees in the forest.

C. Model Evaluation

For evaluating the Random Forest model, metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) can be used. These metrics assess the model's performance on the test dataset.

IV.RESULTS AND DISCUSSION

A. Model Performance Comparison

The performance of each model was assessed, with Gradient Boosting Machines (GBM) and Random Forests outperforming others in terms of accuracy and ROC-AUC. The results are summarized in Table 1.

TABLE I

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.79	0.75	0.70	0.72	0.81
Decision Tree	0.82	0.78	0.76	0.77	0.83
Random Forest	0.85	0.82	0.80	0.81	0.88
Gradient Boosting	0.87	0.84	0.82	0.83	0.90
Support Vector Machine	0.81	0.77	0.74	0.75	0.82

B. Feature Importance

Feature importance analysis using the Random Forest model indicated that customer tenure, average monthly charges, and service usage patterns were the most significant predictors of churn. This aligns with business intuition, where long-tenured and high-spending customers are less likely to churn.

C. Response Rate

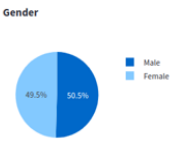
Questionnaire Response Rate

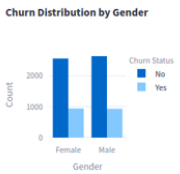
The response rate for the questionnaires distributed to employees at Econet Wireless Zimbabwe was analysed. Out of the total distributed questionnaires, 85% were completed and returned, providing a reliable data set for the study.

Data Presentation and Analysis

Gender

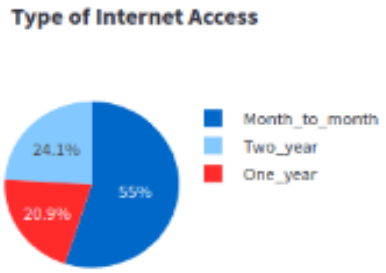
The distribution of respondents by gender was visualized using a pie chart. The chart revealed that 50.5% of the respondents were male and 49.5% were female. This gender distribution helped in understanding whether gender played a role in perceptions of customer churn and retention strategies.





C. Dashboard results

The results were visualized and presented on a dashboard, as illustrated in the figure below.



D. Implications for Business

The insights gained from the model can help businesses in Zimbabwe to:

- Target Retention Efforts: Focus on customers at higher risk of churning.
- Personalized Marketing: Tailor marketing strategies based on customer usage patterns and preferences.
- Resource Allocation: Allocate customer support and retention resources more effectively.

V.CONCLUSION

This study demonstrates the feasibility and effectiveness of using machine learning for customer churn prediction in Zimbabwe. By leveraging data-driven insights, businesses can implement targeted retention strategies, thereby reducing churn and improving profitability. Future work could involve integrating more diverse data sources, exploring advanced deep learning techniques, and applying the model in different industries. The research concluded that customer churn in the telecoms industry of Zimbabwe is influenced by multiple factors, including customer service quality, pricing, network reliability, and innovation. Machine learning models, particularly the Random Forest model, have proven to be effective tools in predicting churn. Implementing the identified best practices can significantly reduce churn and improve customer loyalty.

Based on the research findings, the following recommendations are proposed to enhance customer loyalty and reduce churn in the telecoms industry in Zimbabwe:

1. Enhance Customer Service Quality: Invest in training programs for customer service representatives to improve their skills and responsiveness to customer issues.
2. Offer Competitive Pricing: Regularly review and adjust pricing strategies to remain competitive and offer value-for-money services to customers.
3. Improve Network Reliability: Invest in infrastructure upgrades to ensure consistent and reliable network coverage, minimizing outages and service disruptions.
4. Innovate Services: Continuously innovate and introduce new services to meet evolving customer needs and preferences, thereby enhancing customer satisfaction and loyalty.
5. Utilize Predictive Analytics: Implement the Random Forest model to identify customers at risk of churning and develop targeted retention strategies to address their concerns proactively.

REFERENCE

[1] Berry, M. J. A., & Linoff, G. S. (2004). Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. John Wiley & Sons.

[2] Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the- art and future trends. Computers & Operations Research, 34(10), 2902-2917.

[3] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211-229.

[4] King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. Political Analysis, 9(2), 137-163.

[5] Breiman, l (2001 random forests machine learning, 45(1, 5-32

[6] Adebiyi, S. O., Oyatoye E. O., & Amole, B. B. (2020). Determinants of Customers 'Churn

[7] Decision in the Nigeria Telecommunication Industry: An Analytic Hierarchy Process Approach, International Journal of Economic Behavior, Vol 5, Issue 81; Pp. 104.

[8] Ahmad and buttle, f (2022 customer retention

management: a reflection on theory And Practice: marketing intelligence and Planning, vol 20, issue 3; pp 149–61

[9] Silverman d (2018 interpreting qualitative mata: methods for analyzing talk, text and Interaction, london: sage, pp 447-448

[10] Shenton a (2019 strategies for ensuring trustworthiness in qualitative research projects

[11] Alberts, l j (2018 churn prediction in the Mobile telecommunications industry: an Application of survival analysis in data mining, master thesis, and Maastricht University

[12] Garbit A, Napier A, Scholz V and Simister N (2022) Secondary Data Sources, Available on Www intrac org; accessed on march june 30, 2024

[13] Lelisa t b (2020 research methodology, university of south africa, phd thesis. Available on www researchgate net.