**Group 2 Capstone project proposal**

**Towards Smarter Customer Retention: A Machine Learning Approach to ISP Churn Prediction in Kenya**

**Introduction**

The liberalization and globalization of markets have intensified competition across industries, leading to increased customer turnover. As organizations strive to maintain profitability and sustainability, effectively anticipating and mitigating customer churn has become a critical success factor. Retaining loyal customers is often more cost-efficient than acquiring new ones, yet many firms still struggle to identify and act upon early signals of attrition (Manzoor, Quresh, Kidney, & Longo, 2024). This situation is not an exception when it comes to Internet Service Providers (ISP’s). ISPs, in particular, face this challenge acutely due to the commoditized nature of their services and the relatively low switching barriers for customers.

In Kenya, the ISP landscape has become increasingly competitive in recent years. Established providers such as Safaricom, Zuku, Jamii Telecommunications (Faiba), and Telkom Kenya are now facing stiff competition from emerging players, including regional fiber and wireless service providers entering smaller towns and peri-urban areas. This growth has been driven by rising demand for reliable connectivity to support e-commerce, online education, entertainment streaming, and remote work (ADEMI, 2024). While this expansion benefits consumers, it also heightens the risk of customer churn as users switch to providers offering lower prices, better coverage, or superior customer service.

Customer churn represents more than just lost revenue it directly undermines long term viability in an industry where infrastructure investments are capital intensive and margins can be thin. A high churn rate may indicate dissatisfaction with service quality, pricing models, or support responsiveness, and in turn, erodes brand reputation. According to studies in the telecommunication sector, churn prediction and retention strategies can save firms up to five times the cost of acquiring new customers (Fujo, Subramanian, & Khder, 2022). For Kenyan ISPs, where price sensitivity is high and consumer trust in service providers is often fragile, proactive management of churn risk is not just advantageous but essential.

Machine learning offers a promising pathway for addressing this problem by enabling data driven predictions of churn risk. By analyzing customer behavior, transaction history, usage patterns, and service feedback, predictive models can identify customers likely to leave before they actually churn. Such insights allow ISPs to implement targeted retention strategies, such as offering tailored promotions, improving network quality, or personalizing customer support interventions.

This project builds on that approach, aiming to develop a robust churn prediction model that is specifically tailored to the Kenyan ISP market context.

Ultimately, understanding and predicting customer churn will empower Kenyan ISPs to strengthen customer loyalty, reduce unnecessary marketing expenditures, and allocate resources more effectively. Beyond business performance, successful churn management can also contribute to broader national goals of digital inclusion and economic growth by ensuring more citizens have sustained access to reliable internet connectivity. In this way, the project not only addresses a pressing business problem but also aligns with Kenya’s broader digital transformation agenda under initiatives like the Digital Economy Blueprint (Bunde, 2024)

**Business Understanding**

In this project, churn means when a customer stops using the service or becomes inactive for more than 30 days. For ISPs, this is a serious problem because it causes loss of income and lowers the value of the big investments made in building networks and infrastructure. With more and more ISPs entering the Kenyan market, reducing churn has become an important business goal, not just a technical task.

The project has four main goals:

* Identifying the customers who are most likely to leave by using data and predictive methods.
* Finding the main reasons why customers leave, such as poor service quality, high prices, weak customer support, or changes in usage habits.
* Suggesting simple and effective ways to retain customers that are affordable but also improve satisfaction.
* Providing clear insights from the data that marketing and customer service teams can use to take early action.

By achieving these goals, the project will help Kenyan ISPs lower churn, build stronger customer loyalty, and remain competitive in the fast-changing digital economy. In practice, this means marketing teams can create personalized offers, while customer service teams can get early warnings and act before customers decide to leave.

**Data Understanding**

The study will use a dataset obtained from Hugging Face, which provides a broad view of customer characteristics, behaviors, and feedback, making it well-suited for churn prediction tasks.

The features will be grouped into three main categories:

* Demographic attributes: such as region type (urban, peri-urban, rural) and membership or subscription level.
* Behavioral indicators: including session frequency, activity recency, and last login timestamp.
* Feedback-related measures: such as customer complaints, satisfaction ratings, and sentiment derived from textual feedback.

By capturing these diverse dimensions, the dataset will enable both the prediction of churn risk and the interpretation of the underlying drivers of customer attrition. This is particularly valuable in the ISP context, where customer loyalty is strongly influenced by service quality, pricing models, and overall user experience.

**Data Preparation**

Before building the churn prediction model, the dataset will be cleaned and transformed to ensure it was ready for analysis. The main steps will include:

* Handling missing data by removing rows with gaps in key columns.
* Converting categorical variables (such as region type or membership level) into numerical form using one-hot and ordinal encoding.
* Standardizing numerical features so that values are on a similar scale for modeling.
* Creating new features to add more insights, including:
  + *Customer tenure* (length of time a customer had been with the ISP).
  + *Complaint resolution latency* (how long it took to resolve customer complaints).

These preparation steps will help to improve data quality and ensure the features capture important aspects of customer behavior, service experience, and satisfaction that are relevant to churn prediction.

**Modeling Approach**

In this project, customer churn will be treated as a binary classification problem (a customer will either churn or stay). To address this, five machine learning models will be tested and compared:

1. Logistic Regression – to serve as the baseline model.
2. Random Forest – to provide interpretability and highlight important features.
3. XGBoost – expected to deliver strong performance on structured data.
4. LightGBM – to act as a fast and scalable alternative.
5. MLP Neural Network – to capture more complex, non-linear patterns in customer behavior.

By testing these models, the study will compare performance, evaluate trade-offs between accuracy and interpretability, and select the most suitable approach for predicting churn in the Kenyan ISP market.

**Model Evaluation**

The models will be evaluated using standard classification metrics to ensure both accuracy and fairness in predicting churn. The key metrics will include:

* Accuracy – to measure the overall correctness of predictions.
* Precision – to assess how many of the predicted churn cases are truly churners.
* Recall – to evaluate how well the model identifies actual churners.
* F1 Score – to balance precision and recall for a more reliable measure.
* ROC-AUC – to capture the model’s ability to distinguish between churners and non-churners across different thresholds.

These metrics will guide the selection of the most effective model for churn prediction, with particular emphasis on recall, since identifying customers at risk of leaving is more critical for ISPs than simply achieving high overall accuracy. Based on evidence from prior studies in the telecommunications sector, XGBoost is expected to perform strongly, but its results will be compared fairly against the other models before final selection.

**Deployment**  
The study findings will be shared through clear and practical outputs designed to inform policymakers, labour economists, and other stakeholders:

* **Research Report:**  
  A detailed report will document the methodology, results, and policy recommendations, providing an evidence base for labour market planning and interventions.
* **Interactive Dashboard:**  
  A prototype dashboard (developed in Python Dash or Tableau) will visualize unemployment forecasts alongside official figures. It will allow users to explore trends, compare predicted versus actual unemployment rates, and examine sectoral or regional dynamics.
* **Presentation Slides:**  
  A concise slide deck will be prepared to communicate key findings and recommendations to decision makers in an accessible and actionable format.

**Tools and Methodologies**

**Programming Environment:**  
Python will serve as the primary language, with libraries such as pandas for data manipulation, scikit-learn for machine learning, and statsmodels for econometric and time-series analysis. Development will primarily be carried out in Visual Studio Code (VS Code).

**Visualization Tools:**  
Data exploration and results will be visualized using Matplotlib, Seaborn, and Plotly in Python.

**Modeling Approaches**

The churn prediction framework will treat the problem as a binary classification task (customers either churn or stay). Different machine learning models will be tested, including Logistic Regression for baseline performance, Random Forest for interpretability, Gradient Boosting methods (XGBoost and LightGBM) for strong predictive power, and a Neural Network to capture more complex patterns in customer behavior. Comparing these models will help identify the most effective approach for predicting churn in the Kenyan ISP market.

**Version Control and Reproducibility**

All code, experiments, and results will be tracked using Git and GitHub. This will ensure proper version control, transparency, and collaboration, while also making the research process easy to reproduce in future studies.

**Conclusion**

This study will design and test a churn prediction model tailored to Kenyan ISPs. The insights from the model will help service providers identify customers most at risk of leaving, understand the key reasons behind churn, and apply targeted strategies to improve retention. By doing so, ISPs will be able to reduce customer turnover, strengthen loyalty, and remain competitive in a fast growing and dynamic market.

# Bibliography

ADEMI, D. (2024). *DIGITAL RIGHTS: A CRITICAL ANALYSIS OF THE STATUS OF ONLINE FREEDOM OF EXPRESSION IN KENYA .* Nairobi : Google Scholar .

Bunde, A. O. (2024). The Digital Platform Economy Revolution And Entrepreneurship Ecosystem in Kenya. *SSRN*.

Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). tomer Churn Prediction in Telecommunication Industry Using Deep Learning. *Information Science Letters* , 185-198.

Manzoor, A., Quresh, M. A., Kidney, E., & Longo, L. (2024). A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners. *IEEE Access*, 70434 - 70463.