Domain Background:

Uganda has adopted the use of internet banking, through the use of smart phones, agent banking and through portals that have integrated APIs to transact directly with their customers to streamline access to financial services. In fact, the majority of Uganda's population, both rural and urban population have quickly adopted internet banking as opposed to traditional banking, with over 11 million users on only MTN network from Uganda (Ali et al., 2020; Kafeero, 2020). They use it for savings depositing, microfinance financial management, retail sales, and business-to-business transactions. More specifically, services such as sending and receiving money, paying for utilities, savings schemes, money transfers, and pension payments use internet banking widely in Uganda (Ali et al., 2020). Despite its benefits, there has been a series of concerns in relation to internet banking in Uganda. These include security concerns from identification and validation of account holders, poor network coverage, high transaction costs as reported by certain parts of the population and fraud (Nuwagaba, 2015). To overcome this, there has been an improvement in the uptime of the service providers through scheduled maintenance, the use of biometrics to ensure identity protection and authorisation of users. However, Uganda still experiences a series of challenges in internet banking. One major standout has been fraudulent transactions, leading to heavy losses for users (Nuwagaba, 2015). It has been studied that financial fraud, which severely affects the economy, has led to over $500 billion globally in losses realised by 2020 (Morgan, 2021). In Uganda alone, up to $11 billion was lost in internet fraud in 2019 and a further $20 billion worth of transactions challenged that were bank transfers (Kafeero, 2020).

Therefore, this research proposes using machine learning to overcome fraud challenges in internet banking in Uganda. Studies have shown that machine learning models have previously been adopted to overcome financial fraud, including supervised and unsupervised approaches.

Problem Statement

There is a high level of financial fraud experienced among customers, despite the high adoption of internet banking services in different countries. As of August 2018, up to 90% of financial fraud was experienced in the banking sector globally, with up to 18.28 billion dollars lost in providing financial services through internet banking (Syniavska et al., 2019). Because of this, as of 2019, growth in the use of banking was registered to have been stunted, falling from 3.8% in 2016 growth rate to 2.7 (Magaji, 2020). To overcome this, financial institutions have adopted network security strategies, improved authentication strategies to ensure verification and validation strategies for account holders and made cross-institutional and cross-border banking services available at lower costs. However, there is still a high level of fraud among Uganda's financial institutions. This is likely due to weak identity verification mechanisms, weak intrusion detection systems, crime from within the banking sector or weak coordination mechanisms among banks to detect fraud (Nyakarimi et al., 2020). Therefore, this research proposes developing and adopting a robust machine learning model in online payment systems that use previous banking data from a series of online payment transactions among financial institutions in Uganda to detect possible fraudulent transactions during online payments. This model is viewed as a strategy likely to reduce financial fraud during online payments and increase the general public's confidence in online banking systems.

Solution statement:

The study proposes adopting machine learning models to overcome fraud in online banking through timely detection, based on observed variables of both sending and recipient accounts to correctly classify fraudulent from non-fraudulent transactions (Xiaoli Shen, 2022). Through data mining, machine-learning strategies have been shown to successfully detect financial fraud (Awoyemi et al., 2017). Banking data is predominantly skewed, with the majority of it being non-fraudulent. Therefore, the study proposes using resampling techniques: under-sampling, oversampling, or combining both to generate synthetic samples to train a machine learning model that can enable financial institutions to detect fraud.

Different machine learning models will then be trained, with consideration for the use of grid search. It will be used to identify the best possible parameters for each model. Upon selection of the preferred model, a pipeline will be developed. The researcher proposes using an AWS S3 bucket to store data collected from a weblink as the data source. The Sagemaker pipeline will be developed to retrieve this data from the AWS- S3 bucket, pre-processed, and stored in the feature store. Hyperparameter tuning and training for the model with the data will then be done. An endpoint will then be created, from which the model will be accessible for use through Sagemaker SDK or Amazon services like Lambda and API gateway. A model monitor will be created that will constantly monitor the f1 score of the model, to ensure it is rectified to give absolute predictions on fraudulent transactions, and improved when the performance falls below the recommended scores.

The pipeline will also be designed to sample 40% of data that can be labelled using Sagemaker Ground Truth, which can be used to improve the model frequently. In addition, triggers in the pipeline will be developed at different stages to ensure constant monitoring at different stages. A trigger on the upload of data will be implemented to run the pipeline when and if data is uploaded. Another trigger will be configured at pre-processing. When the code for pre-processing is modified, the pipeline will be configured to run when the git code has been updated, which will then be picked from the git repository and run from the sage maker pre-processing stage to training and inference.

Datasets and Inputs

The study intends to utilise data from 6.3 million banking records to develop, train and evaluate different machine learning models based on identified performance measurement metrics in correctly detecting fraud from the dataset. The data to be used is accessible on Kaggle.com (https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset).

The dataset is comprised of 10 columns, with 6362620, 635447 (99.83%) as non-fraudulent and only 8213 (0.13%) as the fraudulent transactions.

Table 1: Dataset columns and their descriptions

No.	Column name	Description of column
1	step	represents a unit of time where 1 step equals 1 hour
2	type	type of online transaction
3	amount	the amount of the transaction
4	nameOrig	customer starting the transaction
5	oldbalanceOrg	balance before the transaction
6	newbalanceOrig	balance after the transaction
7	nameDest	recipient of the transaction

8	oldbalanceDest	initial balance of recipient before the transaction
9	newbalanceDest	the new balance of recipient after the transaction
10	isFraud	the new balance of recipient after the transaction

Benchmark Model

Different scholars have previously tried to resolve the problem of fraud detection, using a series of algorithms. Valavan & Rita (2023) developed a machine learning model to detect fraudulent transactions using machine learning for financial institutions in India. The scholars used decision trees, random forests, linear regression, and gradient-boosting methods on pre-existing data. They used precision, recall, F1, and ROC as metrics to measure the models' performances (Valavan & Rita, 2023). Perantalu and Bhargav Kiran (2017) in their study for credit card fraud detection utilised both logistic regression and decision trees as classifiers for predictive modelling of credit card fraud detection, utilising information gain to select the attributes and accuracy as the evaluation metric. However, the gap identified is that for data mining problems, especially with class imbalance, f1 score is a preferred metric to recommend for fraud detection problems. Other scholars, Perantalu and Bhargav Kiran (2017), upon using logistic regression and decision trees, discovered that the Decision Trees performed better than Logistic Regression in detecting fraudulent transactions. Shakya (2018) also used Random Forest, Logistic Regression and XGBoost to predict fraud, however, the Random Forest performed significantly better than the two models. He attributed the power of the ensemble method of the random forest in predicting fraudulent transactions. In either case, the scholars used Synthetic Minority Over-Sampling Technique (SMOTE) to solve the class imbalance problem – an oversampling technique.

In this study, the researcher proposes to utilise SMOTE, TOMEK links, and a combination of the two resampling techniques to establish their effect on the performance of the Random Forest, Decision Trees, XG boost and Linear regression models. The researcher also proposes the use of hyperparameter tuning, and grid search, as a means to assess their effectiveness in further improving the performance of the fraud detection models.

EVALUATION METRICS

The researcher proposes the use of the f1 score as a metric to measure the performance of the models. This is a combination of both recall and precision of the selected models in assessing their ability to detect fraudulent transactions. The performance of these models will be compared with those of results from Perantalu and Bhargav Kiran (2017), Shakya (2018) and Perantalu, Valavan & Rita (2023) and Bhargav Kiran (2017).

REFERENCES

Ali, G., Ally Dida, M., & Elikana Sam, A. (2020). Evaluation of key security issues associated with mobile money systems in Uganda. *Information*, *11*(6), 309.

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, 29-31 Oct. 2017). Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI),

Kafeero, S. (2020). *Uganda's banks have been plunged into chaos by a mobile money fraud hack*. Quartz. https://qz.com/africa/1915884/uganda-banks-mtn-airtel-hacked-by-mobile-money-fraudsters

Magaji, B. (2020). A Legal Overview of Electronic Banking in Uganda.

Morgan, R. E. (2021). *Financial fraud in the United States, 2017*. US Department of Justice, Office of Justice Programs, Bureau of Justice ….

Nuwagaba, A. (2015). E-Banking Performance in Uganda: A Case Study of Bank of Uganda. *East Asian Journal of Business Economics (EAJBE)*, *3*(2), 13-20.

Nyakarimi, S. N., Kariuki, S. N., & Kariuki, P. (2020). Risk assessment and fraud prevention in banking sector.

Syniavska, O., Dekhtyar, N., Deyneka, O., Zhukova, T., & Syniavska, O. (2019). Security of e-banking systems: Modelling the process of counteracting e-banking fraud. SHS Web of Conferences,

Valavan, M., & Rita, S. (2023). Predictive-Analysis-based Machine Learning Model for Fraud Detection with Boosting Classifiers. *Computer Systems Science & Engineering*, *45*(1).

Xiaoli Shen, V. J., Xin Huang. (2022). Detect fraudulent transactions using machine learning with Amazon SageMaker.