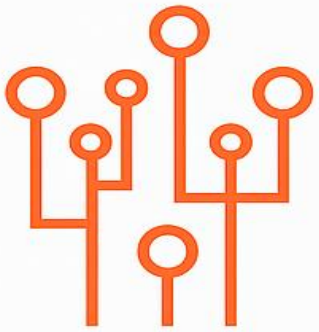

MODELING FIRM-LEVEL LOSS BEHAVIOR AND CORPORATE INCOME TAX (CIT) REVENUE RISK IN KENYA

GROUP 9

1. Introduction

Corporate Income Tax (CIT) remains one of the most fiscally important yet volatile sources of domestic revenue in Kenya. Despite steady growth in nominal economic activity, CIT collections have consistently underperformed targets, contributing to widening fiscal deficits and increased borrowing pressures. A key structural concern underpinning this underperformance is the high prevalence of firms reporting accounting and taxable losses. Preliminary analysis of CIT return data indicates that **approximately 60 percent of firms report losses in a given financial year**, significantly eroding the effective tax base.

While aggregate revenue shortfalls are well documented, less attention has been paid to understanding *why* such a large share of firms consistently report losses, and which firm-level characteristics are most strongly associated with this behavior. This project addresses this gap by using firm-level CIT return data to model loss reporting behavior and quantify its implications for CIT revenue risk.

2. Problem Statement

Kenya has persistently failed to meet CIT revenue targets, even in periods of economic recovery. The coexistence of rising turnover in parts of the corporate sector and widespread loss declarations raises critical questions about the structure of firm costs, capital intensity, related-party exposures, and financial arrangements reflected in tax returns.

Current revenue analysis is largely aggregate and retrospective, limiting the ability of policymakers and tax administrators to proactively identify high-risk firms and sectors. Without firm-level predictive insights, compliance interventions remain reactive, and fiscal planning must contend with significant uncertainty. The central problem this project seeks to solve is the lack of an empirical, data-driven framework for identifying firms most likely to report losses and assessing how this behavior translates into systemic CIT revenue risk.

3. Objectives

The study pursues four interrelated objectives:

1. To empirically identify firm-level financial, structural, and transactional characteristics associated with loss reporting in CIT returns.
2. To develop a supervised predictive model estimating the probability of a firm reporting a loss in a given financial year.
3. To assess the concentration and distribution of loss behavior across sectors, business subtypes, and firm groups.
4. To translate firm-level loss probabilities into insights on aggregate CIT revenue risk and fiscal vulnerability.

4. Research Questions

The analysis is guided by the following questions:

1. Which firm-level variables best explain the likelihood of reporting a loss?
2. How do turnover, cost structures, and capital intensity interact to drive loss outcomes?
3. Do related-party exposures, interest expenses, and royalty payments materially increase loss probability?
4. Which firm segments and sectors account for the largest share of potential CIT revenue erosion?

5. Significance, Stakeholders, and Policy Relevance

This project has direct relevance for multiple stakeholders. For the **Kenya Revenue Authority**, the results support risk-based compliance management by identifying firm profiles associated with persistent losses. For the **National Treasury**, the analysis provides a clearer understanding of structural weaknesses in the CIT base, improving revenue forecasting and fiscal risk assessment.

From a policy standpoint, the findings can inform discussions on capital allowances, financing structures, related-party transactions, and the sustainability of current incentive regimes. By grounding these discussions in firm-level evidence, the project contributes to more credible and targeted tax policy reform in an environment of heightened fiscal pressure.

6. Methodology

6.1 Data Source and Coverage

The analysis uses **firm-level Corporate Income Tax return data from 2017 to 2024**, covering approximately **360,000 firms per financial year**. The unit of observation is the firm-year, allowing for both cross-sectional and dynamic analysis.

6.2 Data Richness and Key Variables

The dataset contains detailed administrative variables, including:

- **Firm identifiers and structure:** PIN_NO, BUSINESS_SUBTYPE, GROUP_, BUSS_REG_DATE, STATION_NAME, COUNTRY_RESI, ENTP_OUTSIDE_KENYA
- **Revenue indicators:** GROSS_TURNOVER, TOTAL_SALES_BUS_PROFF_, GROSS_PROFIT
- **Cost structure:** ODC_TOT_OF_OTHER_DIRECT_COSTS, TOTAL_FACTORY_OVERHEADS, TOTAL_EMPLOYMENT_EXP, FINCEXP_INTEREST_EXP
- **Capital and assets:** TOTAL_FIXED_ASSETS
- **Related-party and cross-border exposure:** SL_DUE_TO_RELATED_PARTIES, UL_DUE_TO_RELATED_PARTIES, OI_ROYALTIES
- **Inventory dynamics:** TOTAL_OPENING_STOCK, TOTAL_PURCHASE_AND_IMPORTS, TOTAL_CLOSING_STOCK
- **Timing and compliance indicators:** FILING_DATE, PERIOD, TRP_FROM_DT, TRP_TO_DT
- **Capital related deductions:** EPZ, SEZ, investment deductions, tax holidays.

This richness enables analysis of operational scale, cost intensity, financing structure, and group-related exposures—key channels through which losses may arise.

7. Modeling Approach

The primary modeling framework is **logistic regression (logit)**, where the target variable is a binary indicator equal to one if a firm reports a loss in a given year. Logit is chosen for its interpretability and suitability for policy-oriented analysis.

While logistic regression serves as the primary benchmark due to its transparency and policy interpretability, the analysis will also explore complementary modeling approaches to test robustness and capture nonlinear firm behavior.

Regularized logistic models (LASSO and Elastic Net) will be used to address multicollinearity and identify the most economically relevant predictors in a high-dimensional setting, while tree-based ensemble methods—specifically Random Forest and Gradient Boosting—will be applied to capture nonlinear relationships and interaction effects between turnover, cost structures, capital intensity, and related-party exposures.

8. Data Preparation

Data preparation involves standardizing firm-year records, handling missing and zero-valued fields, and addressing extreme skewness through winsorization. Financial ratios—such as cost-to-turnover, employment cost intensity, and asset intensity—are constructed to improve comparability across firms. Lagged indicators are generated to capture persistence in loss behavior. Exploratory visualizations include sector-level loss rates, distributions of profit margins, and transition matrices showing movement between profit and loss states over time.

9. Evaluation Strategy

Model performance will be assessed using **AUC-ROC, precision, recall, and F1-score**, with emphasis on identifying high-risk firms rather than overall accuracy alone. Validation follows a time-based split to reflect real-world forecasting conditions.

10. Deployment and Tools

Results will be reported through a structured technical report and executive-ready visual summaries. Where appropriate, a simple dashboard may be developed to demonstrate firm-level risk scoring. The analysis will be conducted in **Python**, using Pandas, NumPy, Matplotlib, Scikit-learn, and Statsmodels. Computation will be performed locally. Due to confidentiality, all data will remain in secure, non-public environments.