MAKERERE  **UNIVERSITY**

DETECTING TAX EVADERS IN UGANDA: A COMPARISON OF LOGISTIC REGRESSION AND FINITE MIXTURE OF LOGISTIC REGRESSION

BY

BRIAN APOLLO TUSUBIRA

BSC (MATH & STAT) (Mak)

A DISSERTATION SUBMITTED TO THE DIRECTORATE OF RESEARCH AND GRADUATE TRAINING IN PARTIAL FULFILLMENT OF THE AWARD OF MASTER OF STATISTICS DEGREE OF MAKERERE UNIVERSITY

JANUARY, 2023

Declaration

This study is original and has not been submitted for other degree award to any other University before.

Sign: _____ Date 11th/01/2023

2018/HD06/13964

Approval

This dissertation has been submitted for examination with the approval of the following supervisors:

Professor Salvatore Ingrassia (PhD)

Department of Economics and Business

University of Catania, Sicily, Italy.

Signature _____ Date 11/01/2023

Dr. Saint Kizito Omala (PhD)

Department of Statistical methods and Actuarial Science,

Makerere University, Kampala, Uganda

Signature _____ Date January 11th, 2023

Dedication

I dedicate this dissertation to the soul of my departed loving mother, Ms. Rebecca Nantongo, for her support and mentorship from childhood. May she continue resting in eternal peace. I also dedicate this work to my brother, Enock Musasizi for the advice and guidance he has given me in the process in my studies.

Acknowledgements

I am extremely delighted to express my deepest appreciation to the “Research Collaborative” team members (Professor Salvatore Ingrassia, Professor Francesca Bassi, Dr. Saint Kizito Omala, Dr. John Bosco Asiimwe and Ms. Zabibu Afazali) who guided me through the development of this dissertation. In a special way, I would like to thank my Lecturer, Mr. Bbosa Francis an Assistant Lecturer at School of Statistics, who provided initial guidance for this research.

I also extend my deepest gratitude to my employer Uganda Revenue Authority, especially the Line-Supervisor, Mrs. Sabah Mohammed Kakooza and Mrs. Blenda Mwogeza for the continuous encouragement they gave me.

I would not have made it without the grace and mercy of God. I am always grateful for the works He has done in my life.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BIC	:	Bayesian Information Criterion
EM	:	Expectation-Maximization
GST	:	Good and Sales Tax
FMLR	:	Finite Mixture of Logistic regression
FY	:	Financial Year
IMF	:	International Monetary Fund
LTO	:	Large Taxpayers Office
MTO	:	Medium Taxpayers Office
OECD	:	Organization for Economic Cooperation and Development
SATR	:	Self-Assessment Tax Regime
UGX	:	Uganda Shillings
URA	:	Uganda Revenue Authority
VAT	:	Value Added Tax
VIF	:	Variance Inflation Factor

Definition of Terms

Assessment:	Is the determination of the amount payable by a taxpayer for a year of income.
C-efficiency:	Is the ratio of VAT revenue to the product of the standard rate and final consumption
Financial Year:	Is a year of income of a taxpayer
Year of income:	Is the period of twelve (12) months ending on June 30th and includes a substituted year of income and a transitional year of income
Substituted year of income:	Is a period of 12 months ending on a date other than June 30th
Transitional year of income:	Is a period of less than 12 months that falls between the person's previous accounting date and a new accounting date
Return:	Is a declaration of the taxpayer's business transaction for a given period.
Tax evader:	A taxpayer who was issued an additional assessment for the declarations made during the Financial Year.
Tax period:	Is one (1) calendar month.

Abstract

This paper presents data mining techniques to predict VAT evaders for inclusion in the audit plan. A dataset of 1,311 registered VAT taxpayers in Large Taxpayers' Office and Medium Taxpayers' Office for the period FY2017/18 across 11 features was used. An exploratory data analysis was used to establish the hidden patterns in the dataset and backward elimination method was used to identify the significant features for model development. Logistic regression (LR) and finite mixture logistic regression with and without concomitant variable were used to detect VAT evaders. BIC was used to select between the FMLR model with and without concomitant variable. Results of each technique were compared and the best technique was chosen based on accuracy, precision and recall were used to evaluate model performance.

Findings of the study showed that number of no sales return, tax office, business sector and number of late payments were identified as significant features in VAT evasion detection. FMLR without concomitant variable had a lower BIC compared to FMLR with concomitant variable and was therefore considered. Model performance evaluation between FMLR without concomitant variable and LR was carried out and FMLR outperformed LR in accuracy, recall and precision. Though LR has been extensively used as a solution to tax evasion problems, the findings of the study suggest that FMLR provide better results compared to LR. The findings of the study can be utilized by URA with emphasis on the four (4) significant variables to detect VAT evaders for inclusion in the Audit plan. URA and future studies may employ other: evader attributes, data mining techniques and model performance evaluation metrics on similar dataset and compare the results.

CHAPTER ONE: INTRODUCTION

1.1 Background

Tax is a mandatory financial charge or levy imposed on a person including a legal entity by the state (Mathews *et al.*, 2018; Jupri & Sarno, 2018). It is one of the most necessary financial resources of a government for accomplishing specific goals (Wu, Ou, Lin & Chang, 2012). Currently, tax authorities use Self-Assessment Tax Regime (SATR)to administer tax, an approach that allows taxpayers to comply with their tax obligations without the interference of a tax official (Uganda Revenue Authority [URA], 2011; Jupri & Sarno, 2018). However, SATR is prone to evasion, in fact Crivelli, Mooij and Keen (2015) and Cobham & Janský (2017) estimate total tax losses at over $400 billion for Organization for Economic Cooperation and Development (OECD) member states annually. They further estimate around $200 billion for lower-income countries. Both studies attribute this loss of revenue to business entities resorting to evading taxes impacting on government income. Tax evasion is an illegal way of avoiding tax or when a person or a business entity intentionally avoids paying his/her true tax liability (Mehta, Mathews, Suryamukhi & San, 2018). Indirect taxes like VAT have greatly been affected by tax evasion. VAT is a levy imposed on the value added to a good or service at each stage of production or distribution (URA, 2013).

In Uganda, VAT contributes to 31.7% of the total revenue and is paid by a person who consumes or imports goods and/or services in Uganda (URA, 2011; 2013). However, IMF researchers (Hutton, Thackray & Wingender, 2014) believe the 28.6% C-efficiency in Uganda is lower than the 48.7% average for Sub-Saharan African countries. From 1991, Uganda Revenue Authority (URA) has been responsible for administering central government taxes by establishing effective and efficient methods to monitor and pursue all cases of tax crime and evasion-related activities (URA, 2011). Other tax-types have had compliance issues but not much affected VAT. By the end of 2005, Uganda's Income tax compliance was as low as 38% (Tusubira & Nkote, 2013).

Developed tax authorities like Directorate General for Taxation (Morocco) and State Revenue Committee (Kazakhstan) in the recent years have setup risk analysis systems for automatic detection of suspicious declarations (Assylbekov, Melnykov & Bekish, 2016; Jihal, Talhaou, Daif & Azzouazi, 2018). Detection of these declarations involved developing classification and

prediction algorithms and models. Due to availability of data from automated systems, researchers, Wu *et al.* (2012) and Assylbekov *et al.* (2016) have improved prediction of tax evasion through data mining techniques. However, Wu *et al.* (2012) transformed numeric variables into categorical variables and Assylbekov *et al.* (2016) used features that were used by the Kazakhstan State Revenue Committee and. González & Velásquez (2013), Roux, Perez, Moreno, Villamil and Figueroa (2018) and Ravisankar, Ravi, Rao and Bose, 2011 also used data mining techniques to detect financial fraud, taxpayers with false invoices, under-reporting tax and return defaulters. Data mining is gaining insights and identifying patterns from data stored in databases in such a way that the patterns and insights are statistically reliable and actionable (Sharma & Panigrahi, 2012). Data mining has played an important role in improving taxpayer compliance since it is able to extract fraudulent behaviors of taxpayers from a large dataset.

1.2 Problem Statement

Selection of suspicious evasion cases for inclusion in the audit plan should be automated to effectively monitor VAT evaders. However, in Uganda this process is manual, that is bases on auditors' judgement or knowledge of taxpayers' behaviour and is hindered by technological bottlenecks and low staff numbers (Almunia, et al., 2017). With an increasing number of taxpayers evading taxes, auditors are overburdened by the manual selection process, delaying completion of audits (Ravisankar *et al*., 2011). Due to delays in audit process, Hutton *et al.* (2014) in their study reveal that VAT compliance gap in Uganda is very large, approximately 60% of potential VAT and 6% of GDP. Unable to effectively close this compliance gap, public investment shall negatively be affected due to the budgetary shortage (Wu *et al*., 2012).

Wu *et al.* (2012) and Jupri and Sarno (2018) applied data mining techniques to develop models that detect potential VAT evaders. The developed models with at least 73% accuracy rate outperformed the baseline models and reduced on time taken by auditors to detect evaders. However, their studies did not apply finite mixture models and focused at only corporate and business entities and income tax. Further to this, there has been limited use of data mining techniques in prediction of VAT evaders in Uganda. The purpose of this study is to utilize data mining methods to develop models that detect VAT evaders in Uganda for inclusion in the audit plan.

1.3 Objectives

1.3.1 General Objective

To develop data mining models that detects Value Added Tax (VAT) evaders in Uganda for inclusion in the tax audit plan.

1.3.2 Specific objectives

a) To identify significant features affecting VAT evasion
b) To develop logistic regression and Finite Mixture of Logistic Regression models that detects VAT evaders
c) To compare performance of the developed models

1.4 Study Hypotheses

a) H_{01}: Taxpayer's tax declaration behavior has no implication in predicting VAT evaders
b) H_{02}: VAT evader status of the taxpayer is not affected by the taxpayer's registration
c) H_{03}: VAT evader status of the taxpayer is not affected by the taxpayer's payment behavior

1.5 Significance of the study

The findings of the study will ease the process of selection of suspicious evasion cases by tax authorities of developing countries. The develops could also be extended to other fraud cases in accounting, banking, insurance and other sectors.

The study will also be referenced by researchers who will intend to carry out research in the same focus area.

1.6 Conceptual framework

The conceptual framework theory was adopted from the review of journals of numerous authors and information acquired from tax experts. These have broken down factors that aid in identifying VAT evaders into three levels: Taxpayer Registration details, Tax declaration and Tax Payment behaviour. Wu *et al*., (2012) believe that registration details of a taxpayer defines the compliance behavior. Mehta *et al.*, (2018) argued that tax declaration behavior by a taxpayer explain the taxpayer behavior while Jupri & Sarno (2018) claim that payment behaviour of a taxpayer explains the compliance behaviour of a taxpayer and defines their ability to evade VAT.

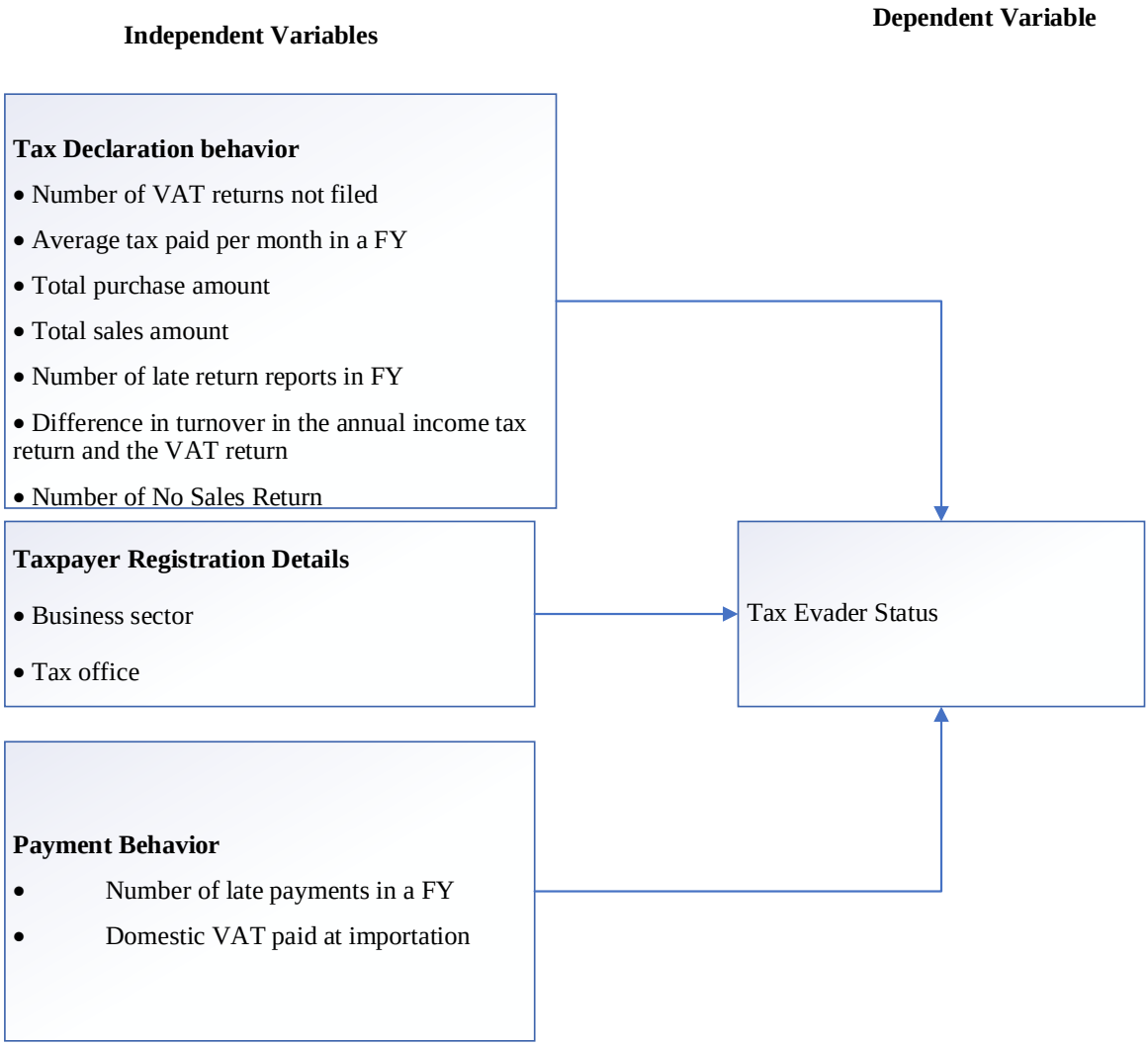


Figure 1.1: Conceptual Framework of VAT evader attributes

1.7 Scope of the Study

The study focused on the features affecting VAT evasion. It was carried out on VAT registered taxpayers from Large and Medium Taxpayers' Offices in Uganda. It considered the Financial Year (FY) 2017/18. The study focused at VAT and particularly FY2017/18 because of the shortfall of the UGX 376.77 billion it registered during the period.

1.8 Organization of the rest of the Report

The rest of this research is organized as follows: Chapter 2 provides the relevant literature that was reviewed. Chapter 3 illustrates the dataset, materials and methods for developing the VAT evasion

models. Chapter 4 presents a discussion of the findings from exploratory data analysis and inferential analysis. Chapter 5 provides the summary of the findings, conclusion, limitations of the study and future studies

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This Chapter provides a review of the available literature on tax evasion, attributes affecting tax evasion and attribute selection. It also reviews data mining techniques for detecting VAT evasion as well as model performance evaluation techniques.

2.2 Definition of tax evasion

Wu *et al.* (2012) in their study considered a tax evader as a taxpayer who was subjected to tax evasion punishments by the tax authority during an audit. They urgue that considering absolute amounts like sales amount may not be good criteria for selecting suspicious evaders. Assylbekov *et al.* (2016) considered the amount of VAT that would be payed by the bussiness entities if they were tax audited. Jupri and Sarno (2018) considered two (2) levels of tax compliance that is formal and material compliance. In their research, they believe an entitity complies formally if it does not report late more than three times in a year and has not paid late more than three times. Material compliance was for taxpayers who have no tax arrears in a year, paid all taxes due and have no difference between their payments and tax declarations. On the other hand, Mawejje and Okumu (2016) measured tax evasion in Uganda using the proportion of sales kept off the books of accounts by firms using data obtained from the World Bank Enterprise Survey.

This study employs Wu *et al.* (2012) and Assylbekov *et al.* (2016) consideration of a tax evader that is a taxpayer that has been issued an additional assessment by URA. Formal compliance by Jupri and Sarno (2018) like no number of late payment may not be good indicators because of the delayed payments especially government suppliers.

2.3 Review on predictors of Tax evasion

Wu *et al.* (2012) in determining a specific approach to improve tax auditor's productivity and performance in handling the detection tasks of VAT evasion in Taiwan, used data samples from 2003 and 2004 collected from the VAT database. Researchers considered features that were organized in three dimensions that is tax evasion control, tax registration and tax files from which

they constructed two datasets (Cubes) having different sets of features to study their dependency. In Cube 1, the features included: the amount of tax evasion, business sector, total capital, sales amount, value-added ratio and sales-capital ratio. In Cube 2, features included; the amount of evasion, business sector, total capital, exemption ratio, sales return ratio and duplicate uniform invoice ratio. The study found out that the accuracy rate of dataset 2 (84.17%) was higher than that of dataset 1 (73.35%).

Assylbekov *et al.* (2016) applied unsupervised learning to detect VAT evasion by business entities in Kazakhstan. They selected 2012 and 2013 business entities which were active in 2011 and 2012 and which had annual income not less than one million Kazakhstani tenge. They developed their model using the 10 features which were used by the State Revenue Committee (SRC) of the Republic of Kazakhstan. These were: coefficient of tax burden, which is the ratio of the assessed amounts of tax revenue without customs payments to the gross annual income before adjustments, the industry average coefficient of tax burden to which a business entity belongs, amount of transactions with false entities, turnover as per VAT and as per the corporate income tax declarations, wage fund, total annual income, assets, de-registration and registration with the tax authorities two or more times a year and whether the Chief Executive Officer (CEO) and/or the founder are declared invalid by the courts. The results of the tax audit were found to be more correlated with the scores given by the developed model than the baseline model of the SRC.

Jupri and Sarno (2018) used features from taxpayer tax return and taxpayer payments of the fiscal year 2014 and 2015 to determine the level of corporate taxpayer compliance using different data mining techniques. These were: number of late return reports in a year, number of non-reporting/filing a year, number of late payments in a year, arrears status, difference in turnover in the annual income tax return with the VAT return, difference turnover where turnover less than the amount withholding tax slip, Regulation Number category, difference in income tax returns with payments and taxpayer compliance status.

Mehta *et al.* (2018) constructed a data set of 58,154 firms with seven features that were; Goods and Services Tax (GST) Identification Number, GST return filing status, number of GST returns not filed by a firm, Geographic division name, ratio which was extracted from the social network, average tax paid per month, total purchase amount and Mean Absolute Deviation (MAD) to develop a predictive model for identifying return defaulters in GST. The developed model reduced

time taken by auditors to identify return defaulters. However, researchers recommend consideration of other tax declarations like income tax declaration to improve the performance of the model.

Mathews *et al.* (2018) in their study to design and implement a regression model that predicts the amount of tax evaded by the potential tax evaders performed a clustering analysis on the Pearson's correlation coefficient between the seven different pairs of tax features. The features in their research included: total sales amount, total GST liability, total state government tax, total state government tax paid in cash, total exempt sales, total liability, total Input Tax Credit and Integrated GST. Each feature considered was a vector of 12 values relating to the tax return information of 12 months in the Financial Year (FY) 2017/18. Using the developed model, researchers discovered that the percentage of total amount of tax evaded by suspicious set of dealers was 81.43%.

From the above studies, this study applied the following features; number of VAT returns not filed, average tax paid per month during a certain tax period, total purchase amount, number of late return reports in a year in a year, number of late payments in a year, difference in turnover in the annual income tax return with the VAT return, business sector and total sales amount paid by the taxpayer during the FY. It also considered; number of no sales returns, tax office and domestic VAT paid at importation following engagements with URA tax auditors on factors that influence VAT evasion.

2.4 Data mining techniques for detection of VAT evaders

Data analysts have used various data mining techniques to detect tax evasion that include: Decision trees, Neural Networks (NN), Association rule, Bayesian networks (BN), Logistic Regression (LR) and Support Vector Machines (SVM).

Ravisankar *et al.* (2011) used data mining techniques such as Multilayer Feed Forward Neural Network (MLFF), SVM, Genetic Programming, Group Method of Data Handling, Logistic Regression and Probabilistic Neural Network (PNN) to predict the occurrence of financial statement fraud in 202 Chinese companies. From 35 features, researchers used the t-statistic test feature selection method to develop classification algorithms using the top 18 features and later

using the top 10 features. Results highlighted that PNN outperformed all the other techniques followed by GP.

Wu *et al.* (2012) applied association rule to identify and select suspicious VAT evasion reports for further auditing in Taiwan. In selection of the data sample and data preprocessing, researchers used Visual Basic scripts on Standard Query Language Server 7.0. In addition, the attribute domain used consisted of numeric data that were transformed into categorical data which impacted on its efficiency. Results showed that the proposed data mining technique enhanced the detection of tax evasion and can be employed to effectively reduce losses from VAT evasion compared to the manual screening method. Researcher advised that using the data mining technique on a large amount of tax data can improve the accuracy rates in screening potential tax evasion reports.

González & Velásquez (2013) in their comparative study of data mining techniques to characterize and detect potential users of false invoices in a given year used clustering algorithms of Self Organizing Maps (SOM) and Neural gas to identify taxpayer groups of similar behavior. Decision trees, Neural Networks and Bayesian networks were used to identify variables that were related to conduct of fraud and/or no fraud, detect patterns of associated behavior and establishing to what extent cases of fraud and/or no fraud can be detected with the available information. Though the results showed that Neural Networks were the best detection data mining technique with 92.6% of fraud cases assigned to correct classes and followed by Decision trees with 89.0%. However, the researchers only considered one attribute of tax evasion "false invoice".

Assylbekov *et al.* (2016) in their study used unsupervised model of SOM to detect VAT evasion by business entities in Kazakhstan. Using Kohonen package for R, researchers trained SOMs separately for each of the considered year (2012 & 2013) and discovered that the results of tax audit were more correlated with scores given by the developed model than with the scores by the baseline model used by SRC. However, the developed model based on an assumption that features of a tax compliant business entity follows a multivariate Gaussian distribution $N(\mu, \varepsilon)$ which is not always the case.

Egger, Merlo and Wamser (2018) used finite mixture modeling approach to investigate the tax responsiveness of multinational firms' investment decisions in foreign countries classifying them into avoiders of profit taxes and those that are not. They discovered that a one-percent- increase in

the statutory corporate profit tax rate of a host country is found to reduce the fixed assets of non-avoiders in that host country by 0.81%.

Mathews *et al.* (2018) applied various data mining algorithms to develop a technique that predicts the amount of tax revenue lost by the state due to deceitful actions and classify dealers as genuine or suspicious. Clustering algorithm was used to cluster dealers, Benford's analysis to classify genuine and suspicious dealers and the linear regression to develop suspicious and genuine dealer models. The built linear regression model was used to predict the amount of tax evaded by potential tax evaders and gave an adjusted R-squared value of approximately 94.0%.

Jupri and Sarno (2018) in their study to detect and classify the level of corporate taxpayer compliance in Indonesia used various classification algorithms C4.5, SVM, KNN, Naïve Bayes and MLP. The classification results of each algorithm were compared and the decision tree C4.5 algorithm had the highest preference value basing on F-score, Accuracy and time taken to build the model compared to other algorithms. However, study was only limited to VAT compliance of co-operate taxpayers.

Mehta *et al.* (2018) in their paper constructed a logistic regression model that predicts whether a business entity is a potential return defaulter for the upcoming GST filing period. Researchers computed the MAD value to test if the data's first digits conform to the expected probability distribution from which they detected fraud using Benford's law.

Roux *et al.* (2018) presented a novel approach for the detection of potential fraudulent taxpayers using only unsupervised learning techniques. The study aimed at developing a screening technique for detection of under-reporting tax declarations in Bogota, Colombia without having historic labeled data. Researchers used a three-fold process that involved clustering to ensure tax declarations are grouped according to the values of their features, adjusting the probability distribution and finally detecting suspicious declarations using a quantile of the adjusted distribution. From the developed model, none of the non-under-reporting declarations were marked by the auditors as suspicious and only one of the under-reporting declarations were marked as unsuspicious by the auditors. However, the study was limited to only the construction sector in due to the high risk that builders under report their budget in order to avoid paying a higher tax base.

The dataset provided was labelled, developing models on training dataset with labelled class would give better performance compared to KNN, PNN, SOM methods that do not need labelled data. Decision trees and association rule are more prone to over-fitting. Logistic regression on the other hand is less inclined to over-fitting and finite mixture models take consideration of unidentified heterogeneity in the dataset. Therefore, the research employed logistic regression and finite mixture of logistic regression with and without concomitant variable.

2.5 Model Comparison

In a review to identify companies that resort to financial statement fraud, Ravisankar *et al.* (2011) presented the average accuracies, sensitivities, specificities and Area Under the receiver operating characteristic Curve (AUC) for the test data over 10-folds. They ranked classifiers basing on Area Under the receiver operating characteristic Curve (AUC) an evaluation metric for binary classification problems. Researchers observed that PNN was the best classifier among all others classifiers in terms of AUC, accuracy, sensitivity and specificity.

In Mathews *et al.* (2018), model performance was determined using adjusted-R^2 value and Root Mean Square Error (RMSE). The adjusted-R^2 value of 0.937 implied that the model was not under fitting while RMSE value on test and train data of 0.000411 indicated that it was not over fitting.

Jupri and Sarno (2018) used Fuzzy TOPSIS method to measure the performance of the classification algorithms. The research weighted the value of F Score, Accuracy and Time taken to build the model to determine the ranking. They established that C4.5 algorithm outperformed the other algorithms.

Mehta *et al.* (2018) used confusion matrix, ROC curve, log likelihood Chi-square test and lift chart to evaluate the performance of the built model. Results showed that the values of the area under the training ROC curve and the testing ROC curve was almost the same indicating that the model was not over fitting. Results from the Confusion matrix showed that the built model achieved a prediction accuracy of 87.0%.

Lahann, Scheid and Fettke (2019) in their study to reveal VAT compliance violations in accounting data using machine learning techniques evaluated the performance of the built models using confusion matrix, Mathews Correlation Coefficient (MEC) and AUC. The Confusion matrix was used to compute the average accuracy, mean average precision and mean average recall. The

C4.5 decision tree performed better than other models with accuracy, precision, recall, MEC and AUC of 0.987, 0.989, 0.939, 0.978 and 0.986 respectively.

There is need to assess the models' ability to give correct predictions thus accuracy shall be vital in the study. However, this case study being an imbalanced class problem with the rate of having an evader being low, accuracy may not be a conclusive measure to assess model performance. In addition to accuracy, focus shall be put on evaluating the model ability to find evader cases with in the data that is recall and precision as applied by Mehta *et al.* (2018) and Lahann, Scheid and Fettke (2019). It is from these considerations that this study shall employ accuracy, recall and precision to compare the performance of the models.

CHAPTER THREE:METHODOLOGY

3.1 Introduction

This Chapter gives a description of research methods that were followed in the study. It provides information on the study population, methods that were used in this study and the justification for using them.

3.2 Study Population

The study population used was composed of 1,311 registered VAT Large and Medium taxpayers in Uganda for the FY2017/18. Only FY2017/18 dataset was considered to minimize analysis errors that could result due to economic policies and other factors introduced in other FYs.

Large taxpayers are taxpayers whose asset base exceeds 30 billion shillings and Medium taxpayers are those whose asset base exceeds 15 billion shillings but less than 30 billion shillings. Taxpayers in these categories contributed to 68.30% of URA's total domestic revenue collection that FY (URA, 2018).

3.3 Data collection

Dataset used in this research was obtained from the URA oracle database and contained 1,311 VAT taxpayers for the FY2017/18. 448 of these were from Large Taxpayer's office (LTO) and the rest were from Medium Taxpayer's Office (MTO).

For each taxpayer, 12 features were obtained from the taxpayers' Registration details, tax declaration behavior, tax payment behavior and tax evasion status as shown in Table 3.1

Table 3.1: Description of data attributes

SN	Attribute Name	Description	Data type
1	Business Sector	Business sector for a taxpayer.	Categorical
2	Tax Office	Tax office of the taxpayer.	Categorical

SN	Attribute Name	Description	Data type
3	RNF	Number of VAT returns not filed by a taxpayer in a Financial Year.	Numeric
4	NLR	Number of late return reports in a Financial Year	Numeric
5	TSA	Total sales amount declared by a taxpayer declared in a certain Financial Year.	Numeric
6	TPA	Total purchase amount declared by a taxpayer in a certain Financial Year.	Numeric
7	DIVR	Difference in turnover in the annual income tax return with the VAT return.	Numeric
8	ATPM	Average VAT tax paid per month during a certain Financial Year.	Numeric
9	NLP	Number of late payments in a Financial Year.	Numeric
10	NONS	Number of no Sales return in a Financial Year	Numeric
11	DVAT	Domestic VAT paid at importation in a Financial Year	Numeric
12	Tax Evader Status	A taxpayer who was issued an additional assessment or not	Categorical

3.4 Data Management Process

The researcher applied Knowledge Discovery in Databases (KDD) process for data management. KDD is the nontrivial extraction of implicit, previously unknown and potentially useful knowledge from data and involves data selection, data preprocessing, cleaning, transformation, data mining/analysis and evaluation (Verma, 2015).

3.4.1 Data selection, pre-processing, cleaning and Transformation

Data was exported from the URA oracle database and saved as a CSV file. In development of the model, irrelevant data like tax agent's name and return received time were removed from the selected data. As part of the data cleaning process, taxpayers who amended the tax returns, the last

amended return was considered. Numeric features: TSA, TPA and ATPM were log-transformed to ease interpretation of their patterns.

3.4.2 Data Analysis and Evaluation

The CSV data was imported into R for analysis. R is flexible and has no cost (Mehta *et al.*, 2018). Data analysis involved exploratory data analysis to understand the data, development of the models and evaluation of their performance. Exploratory data analysis involved conditioning each feature to tax evasion status to identify the tax evasion behavior for each feature. Data analysis involved splitting the dataset into 80% training and 20% testing datasets as applied by Shadi, Monther and Muneer (2018) and Ha *et al*. (2018). The former was used to train the logistic regression model and the latter was used to compare the effectiveness of the developed models.

3.5 Significant feature selection

In order to increase prediction performance of the models, feature selection was carried out to remove irrelevant and redundant features (Ramya & Kumaresan, 2015*;* Qasim & Algamal, 2018). Exploratory Data Analysis involving graphs of conditional distribution of each feature on VAT evasion status was carried as a preliminary stage to identify potential features that affect VAT evasion status. A number of studies provide an overview of feature selection methods: this study referred to Unler and Murat (2010) approach during which the researchers used the stepwise backward elimination method using Logistic Regression Equation.

Stepwise backward elimination begins with the full set of features and progressively removes one or more features from the set (Tang, Alelyani & Liu, 2014; Jovic, Brkic & Bogunovic, 2015). Features that are found to have a ρ-value less than (α=0.05) were then entered into a binary logistic regression and finite mixture of logistic regression models (Ranganathan, Pramesh and Aggarwal, 2017)

3.6 Developing the data mining models

The significant features identified were used to develop Logistic regression and Finite Mixture of Logistic Regression models that detects VAT evaders.

3.6.1 Logistic regression (LR)

In this study, LR was considered since it is a well-established statistical method for predicting binary outcomes as recommended by Unler & Murat (2010) and Sahin & Duman (2011). In comparison to other sigmoid models, the fit of Logistic model shows better indicator values (Luigidell'Olio, AngelIbeas, Juan deOña & Rocio deOña, 2018)

The Logistic Regression model is given by Equation 3.1

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \quad (3.1)$$

And the logit-odds Equation given by:

$$(\frac{p}{1-p}) = \ln(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}) \quad (3.2)$$

Where:

p is the probability of a taxpayer being a VAT evader

β_k is the regression coefficient

x_{ki} is the value or category of the i^{th}taxpayer for the k^{th}feature under consideration.

3.6.2 Finite Mixture of Binomial Logistic Regression (FMLR)

Finite mixture models are flexible approach to statistical modeling of a wide variety of data characterized by unobserved heterogeneity (McLachlan & Peel, 2000). The study used finite mixture models to develop an algorithm that detects the different taxpayer behavior that is VAT evader and non-evader. This involved consideration of the model with and without concomitant variable and the model with the lower BIC value was considered. Ignoring concomitant variables can lead to skewed data (Grün & Leisch, 2007).

Model with Concomitant variable

Let X denote the significant features, W the concomitant feature and Y the VAT evasion status and N the samples with each sample having $M_n$observations (x_i, y_i) $(i = 1, \dots, M_n, n = 1, \dots, N)$.

The FMLR model with concomitant variable is given by;

$$h(y|x,w,\varphi(\pi_i,\dots,\theta_i)) = \sum_{k=1}^{K} \pi_k(w,\alpha) f(y|x,\tilde{\theta}_k) \qquad (3.3)$$

Where: $\sum_{k=1}^{K} \pi_k = 1,\ \pi_k > 0$

$h(y|x,w,\varphi(\pi_i,\dots,\theta_i))$ is the conditional density of y given x, w and $\varphi(\pi_i,\dots,\theta_i)$

$\pi_k(w,\alpha)$ is the mixing proportion/ prior probability that a taxpayer y_i belongs to a component k

$\tilde{\theta}_k$ is the component-specific parameter vector for the density function f

$\varphi(\pi_i,\dots,\theta_i)$ is the vector of all parameters

Model without Concomitant variable

Let X denoted the significant features and Y the VAT evasion status and N the samples with each sample having $M_n$observations (x_i, y_i) $(i = 1,\dots,M_n, n = 1,\dots,N)$.

The general FMLR with K components has the form:

$$h(y|x,\varphi(\pi_i,\dots,\theta_i)) = \sum_{k=1}^{K} \pi_k f(y|x,\tilde{\theta}_k) \qquad (3.4)$$

Where: $\sum_{k=1}^{K} \pi_k = 1, \pi_k > 0$

$h(y|x,\varphi(\pi_i,\dots,\theta_i))$ is the conditional density of y given x and $\varphi(\pi_i,\dots,\theta_i)$

π_k is the mixing proportion/ prior probability that a taxpayer y_i belongs to a component k

$\tilde{\theta}_k$ is the component-specific parameter vector for the density function f

$\varphi(\pi_i,\dots,\theta_i)$ is the vector of all parameters

Parameter of FMLR can be estimated through the EM algorithm (Grün & Leisch, 2007). With the E-step estimating the conditional component probabilities (posterior probabilities) for each observation using:

$$\Pr[y_i \in k|x_i, \theta_k] = \frac{\pi_k f_k(y_i|x_i\theta_k)}{\sum_{j=1}^{K} \pi_j f_j(y_i|x_i\theta_j)} \quad (3.5)$$

Where: Θ_k is the vector of component (k)specific parameters.

π_k is the mixing proportion for the k^{th} component

And the M-step involves maximizing the log-likelihood for each component separately using the conditional probabilities as weights.

The E and M steps are repeated until conversion takes place (Grün & Leisch, 2007)

3.7 Performance evaluation of the models

With respect to the existing literature, BIC, Accuracy, Recall and Precision were used to compare the performance of the models. BIC was used to select between the FMLR model with and without concomitant variable and select the most appropriate components to be considered in the FMLR model. The model with the lower BIC is preferred. A confusion matrix was used to compute Accuracy, Precision and Recall. The model with the highest accuracy, precision and recall is better in comparison to the one with a lower value. A confusion matrix allows visualization of the performance of classification model on a set of test data for which the true values are known (labelled data).

Table 3.2: Confusion matrix for model comparison

		Predicted	
		Non-Evader	***Evader***
Actual	***Non-Evader***	True Non-Evaders (TN)	False Evaders (FE)
	Evader	False Non-Evaders (FN)	True Evaders (TE)

Where: $TN = True\ Non-Evaders, TE = True\ Evaders, FE = False\ Evaders, FN = False\ Non-Evaders$

Accuracy (AC) is the proportion of correctly detected taxpayer behavior. It was determined by Equation 3.5:

$$AC = \frac{TE+TN}{(TN+FE+FN+TE)} \quad (3.6)$$

Recall or True Positive rate (TP) is the proportion of actual evasion cases that were correctly classified by the model, calculated using Equation 3.6:

$$Recall\ (TP) = \frac{TE}{FN+TE} \quad (3.7)$$

Precision (P) is the out of all cases that were predicted as evaders, how many were correctly classified by the model. It was calculated using the Equation 3.7:

$$Precision\ (P) = \frac{TE}{TE + FE} \quad (3.8)$$

The model with a higher accuracy, recall and precision is preferred Lahann et al. (2019).

3.8 Ethical consideration

Due to sensitivity and confidentiality of URA data, I sent a request to collect data using my URA work mail to URA Assistant Commissioner Human Resource (AC-HR). The request was granted and an offer letter was provided on 3rd January 2020 and an electronic oath of secrecy and confidentiality was sent. Later, I submitted Makerere University data collection request to AC-HR. Both the data collection request from Makerere University and URA's approval for data collection are attached (Appendix 1 and Appendix 2)

CHAPTER FOUR: PRESENTATION AND DISCUSSIONS OF FINDINGS

4.1 Introduction

This Chapter contains Sections of the data analysis and results of the research that was conducted. It involves exploratory data analysis and inferential analysis for attribute selection, Logistic and finite mixture of logistic regression models. It consists of evaluation of model performance.

4.2 Description of variables

The data studied in this research consisted of 1,311 taxpayer records. Majority of the taxpayers were classified as non-evaders (71.3%). The results further highlight that most taxpayers were from MTO tax office (65.8%). G-Wholesale & Retail Trade; Repair of Motor Vehicles & Motorcycles at 44.6% had the highest number of taxpayers among the 20 business sectors considered (Table 4.3). The largest proportion (99.3%) of taxpayers did not pay Domestic VAT at importation.

Table 4.1: Description of categorical variables

Variable	Categories	Frequency	Percentage (%)
Evader Status	Evader	376	28.7%
	Non-Evader	935	71.3%
Tax Office	LTO	448	34.2%
	MTO	863	65.8%
DVAT	No	1302	99.3%
	Yes	9	0.7%

From Table 4.2, the number of no sales returns ranged from 0 to 12 with an average of 1. Averagely no return either not filed or was filed late while averagely taxpayers made 2 late payments.

Total sales amount per taxpayer ranged from 0 Uganda shillings to 1,943,820,000,000 UGX centering at around 50,758,747,596 UGX. The average of total purchase amount per taxpayer was 6,759,080,557 UGX while the average of average tax paid per month is 19,479,163 UGX. Difference in income tax and VAT averaged at -1,547,802,800 UGX and ranged from -1,019,450,000,000 UGX to 1,622,890,000,000 UGX.

Table 4.2: Description of numerical variables

Variable	Median	Mean	Minimum	Maximum
NONS	0.0000	0.9695	0.0000	12.0000
NLR	0.0000	0.2121	0.0000	10.0000
RNF	0.0000	0.0038	0.0000	1.0000
NLP	1.0000	2.2570	0.0000	12.0000
TSA	15,748,098,261	50,758,747,596	0.0000	1,943,820,000,000
TPA	1,585,204,397	6,759,080,557	0.0000	1,380,360,000,000
DIVR	1,166,435	-1,547,802,800	-1,019,450,000,000	1,622,890,000,000
ATPM	4,381,519	19,479,163	0.0000	1,534,854,914

4.3 Data Extraction

1,311 registered LTO and MTO VAT taxpayers' information for FY2017/18 was extracted from an oracle database of URA. Evasion status with two responses "Evader" and "Non-evader" was the class label with 71.3% of the taxpayers were VAT Non-evaders (Figure 4.1)

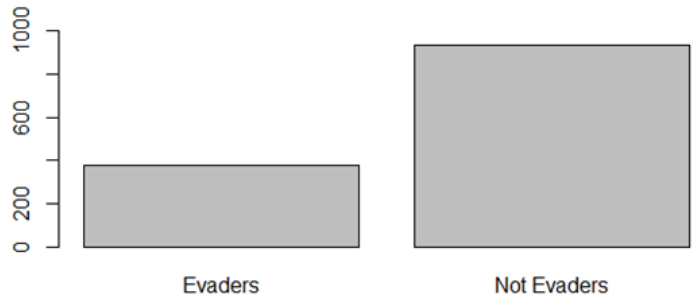


Figure 4.1: Class label distribution

4.4 Data Pre-processing

Data pre-processing involved identifying missing values, converting numeric variables (TSA, TPA, DIVR and ATPM) that were exported as character variables. The data provided had no missing values. Further to that, Business sector variable initially had 20 levels and these were merged to 6 levels (Table 4.3) basing on similarity in operation and governance.

Table 4.3: Merged Business Sectors

Considered Business Sectors	Original Business Sectors	Taxpayers per Sector	Considered taxpayers
A	A - Agriculture, Forestry & Fishing	40	83
	I - Accommodation & Food Service Activities	43	
C	C – Manufacturing	237	272
	D - Electricity, Gas, Steam & Air Conditioning Supply	30	
	E - Water Supply; Sewerage, Waste Management & Remediation Activities	5	
F	F – Construction	115	155
	L - Real Estate Activities	40	
G	G - Wholesale & Retail Trade; Repair of Motor Vehicles & Motorcycles	585	585
H	H - Transportation & Storage	57	102
	J - Information & Communication	45	
O	M - Professional, Scientific & Technical Activities	37	
	B - Mining & Quarrying	11	114
	K - Financial & Insurance Activities	14	
	N - Administrative & Support Service Activities	27	
	O - Public Administration & Defense; Compulsory Social Security	3	
	P – Education	1	
	Q - Human Health & Social Work Activities	7	
	R - Arts, Entertainment & Recreation	4	
	S - Other Service Activities	9	
	U - Activities of Extraterritorial Organizations & Bodies	1	
Total			1,311

4.5 Data transformation

The extracted dataset was converted to a form appropriate for the data mining objective. Continuous variables (TSA, TPA and ATPM) were log transformed to have their distributions less skewed and have easily interpretable patterns.

4.6 Exploratory Data Analysis

A. Analysis of univariate statistics

a) Categorical variables

From the bar-graph below, Tax office, DVAT, Evader Status and RNF have two (2) levels each while Business sector has six (6) levels.

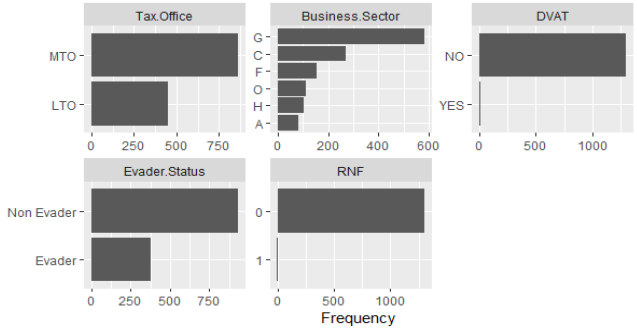


Figure 4.2: Categorical variable bar graphs

b) Numeric variables

From the density plots of numeric variables below, numeric variables (ATPM, TPA and TSA) displayed a left-skewed distribution after transformation (Figure 4.4).

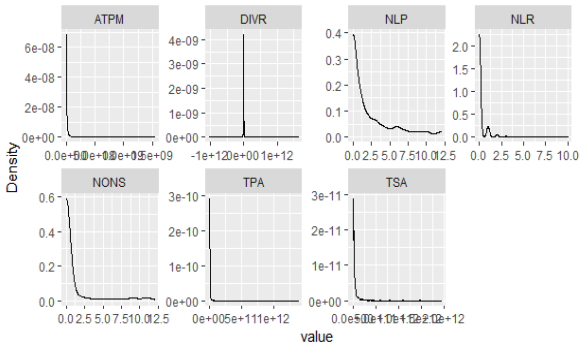


Figure 4.3: Density plots before transformation

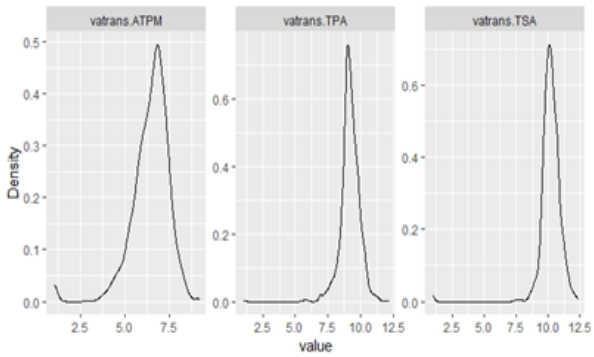


Figure 4.4: Density plots after transformation

B. Bivariate analysis of numeric variables

An analysis of the linear relationship between continuous variables was carried out and below is the correlation matrix:

	TSA	TPA	DIVR	ATPM
TSA	1.00000	0.35587	-0.31276	0.59282
TPA	0.35587	1.00000	-0.06073	0.34724
DIVR	-0.31276	-0.06073	1.00000	-0.02738
ATPM	0.59282	0.34724	-0.02738	1.00000

From the correlation matrix above, there is a moderately positively correlation between TSA and ATPM, a weak positive correlation between TSA and TPA. A feasible explanation is that as a taxpayers' sales increase his or her ATPM is expected to increase as well as the business inputs (purchases). There is a weak negative correlation between TSA and DIVR while the other features are nearly not correlated.

C. Conditional distribution analysis on VAT Evader Status

From the conditional graph, it is more likely to have a VAT evader in LTO than MTO but not big enough to show a difference (Figure 4.5). This could be as a result of relaxation in the monitoring and implementation of compliance initiatives like spot check and comprehensive audits on LTO taxpayers by the tax authority. Taxpayers having more than 5 late returns or at least 1 return not filed are more likely to be VAT evaders as evidenced in Figure 4.6 and Figure 4.8 respectively. Exploratory data analysis displays that VAT evasion status of a taxpayer is not determined by the number of late payments and their respective business sector (Figure 4.7 and Figure 4.9). Analysis also shows that the larger the TSA, TPA and ATPM the more probable the taxpayer is a VAT evader (Figure 4.13 to Figure 4.18). This agrees with taxation practice since as a business expands its transactions become large in numbers and accountants under report sales which are identified as evasion cases. Figure 4.12 further indicates that a taxpayer having more than 10 no sales return is a signal for possible VAT evasion. However, there is no change in compliance behavior of taxpayers who have a difference between their Income tax and VAT sales declaration (DIVR) or those that did not pay Domestic VAT at importation (DVAT) (Figure 4.10 and Figure 4.11).

From the conditional exploratory data analysis: RNF, NLR, TSA, TPA, ATPM and NONS could be important features for the development of model. Since graphics are data detection methods that give clues about the data (Tukey, 1977), there is need to carry out inferential data analysis about the significance of these feature to VAT evasion detection.

Figure 4.5: Tax Office Vs VAT Evader status

Figure 4.6: Number of late returns filed Vs VAT Evader status

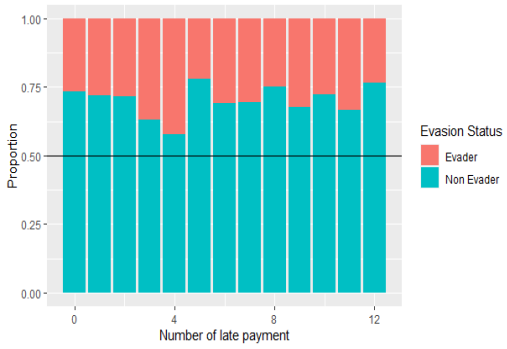


Figure 4.7: Number of late payments Vs VAT Evader status

Figure 4.8: Number of returns not filed Vs VAT Evader status

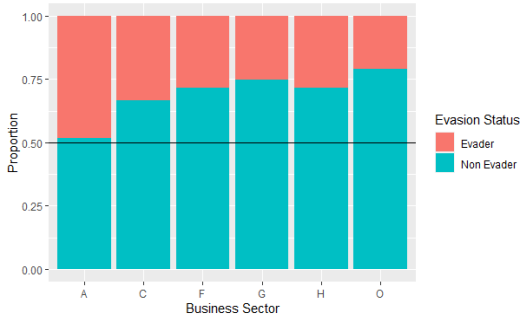


Figure 4.9: Business Sector Vs VAT Evader status

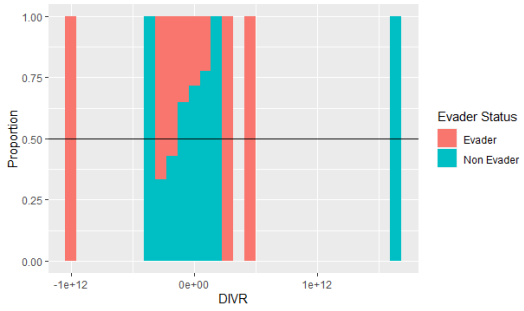


Figure 4.10: Difference in VAT and Income tax turnover Vs VAT Evader status

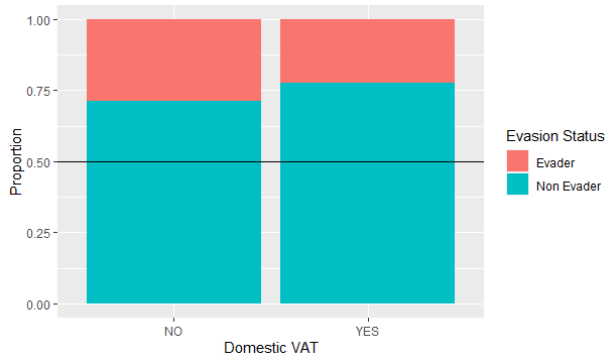


Figure 4.11: Domestic VAT paid Vs VAT Evader status

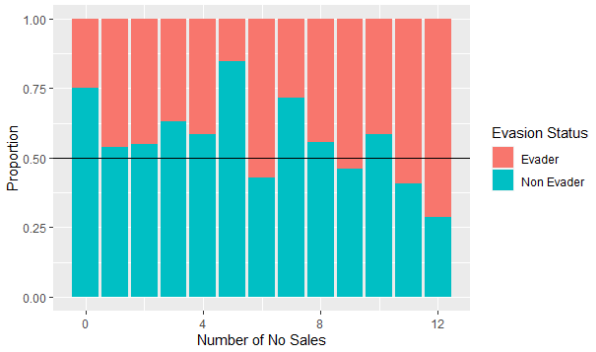


Figure 4.12: Number of no sales returns Vs VAT Evader status

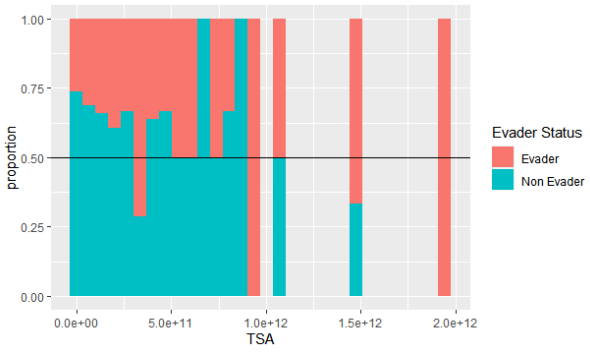


Figure 4.14: Total Sales Amount Vs VAT Evader status before transformation

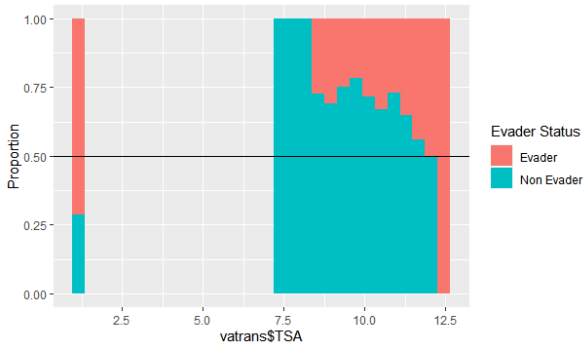


Figure 4.13: Total Sales Amount Vs VAT Evader status after transformation

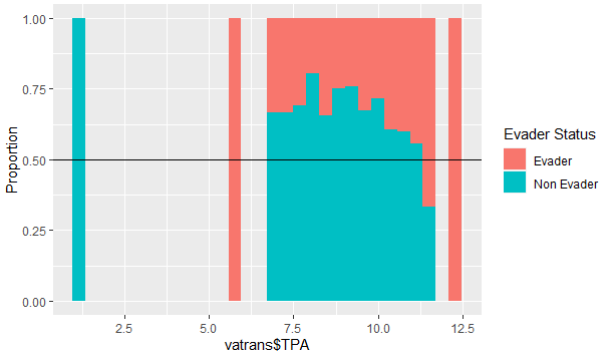


Figure 4.16: Total Purchases Amount Vs VAT Evader status after transformation

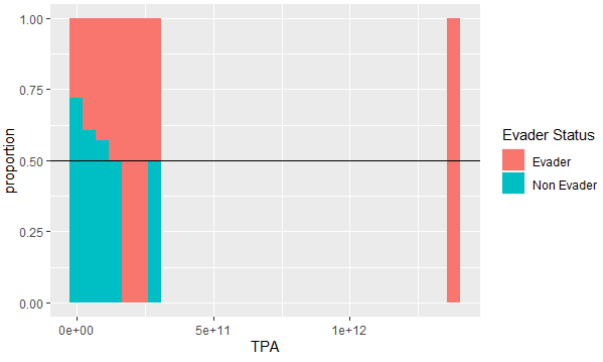


Figure 4.15: Total Purchases Amount Vs VAT Evader status before transformation

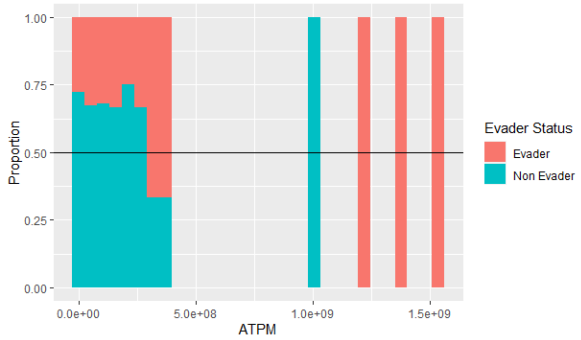


Figure 4.17: Average tax paid per month Vs VAT Evader status before transformation

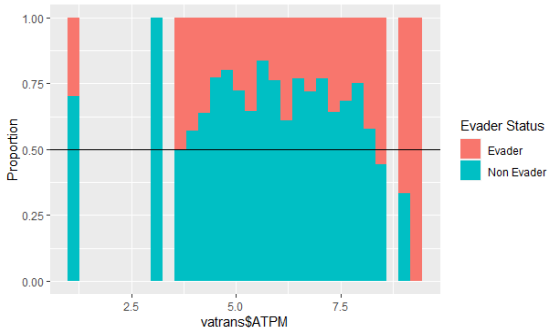


Figure 4.18: Average tax paid per month Vs VAT Evader status after transformation

4.7 Inferential Data Analysis

Though researchers (Wu et al., 2012) and (Assylbekov et al., 2016) used different financial year datasets as training and testing datasets their models, this study only considered FY2017/18 dataset to avoid errors that might arise due to other factors that may have happened in one of the financial years under consideration. The dataset was randomly split into 80% training dataset and 20% testing dataset.

The training dataset consisted of 72.68% non-VAT evaders and thus was up-sampled. Up-sampling was carried to ensure that the dataset has equal proportion of classes to avoid data imbalance problems such as bias for majority class and ignore the minority class (Ali, Salleh, Saedudin, Hussain & Mushtaq, 2019).

4.7.1 Significant feature for model development

Data mining involved identifying the right features in order to develop models that detect VAT evaders. A stepwise backward elimination method for Logistic regression was used. A Logistic regression (Equation 3.1) was applied to the up-sampled training dataset with all variables and the following results were obtained:

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0963	-1.0764	-0.3423	1.1752	1.6039

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.540523	0.946748	0.571	0.56805
Tax.OfficeMTO	-0.411942	0.13495	-3.053	0.002269 **
Business.SectorC	-0.619888	0.236049	-2.626	0.008637 **
Business.SectorF	-1.164902	0.260614	-4.47	0.000008 ***
Business.SectorG	-0.816084	0.222883	-3.661	0.000251 ***
Business.SectorH	-1.013937	0.2814	-3.603	0.000314 ***
Business.SectorO	-1.283287	0.27919	-4.596	0.000004 ***
DVATYES	-0.115603	0.657892	-0.176	0.860516
NONS	0.138362	0.024165	5.726	0.000000 ***
NLR	0.146919	0.082986	1.77	0.076659
RNF	14.358171	341.52178	0.042	0.966465
TSA	-0.007594	0.069679	-0.109	0.913211
TPA	0.014335	0.104667	0.137	0.891064
DIVR	0.000000	0.000000	-0.439	0.660756
ATPM	0.027092	0.054253	0.499	0.617517
NLP	0.0475	0.017783	2.671	0.007560

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2112.7 on 1523 degrees of freedom
Residual deviance: 1996.7 on 1508 degrees of freedom
AIC: 2028.7

Number of Fisher Scoring iterations: 13

At 95% confidence level, inferential analysis shows that ATPM, RNF, NLR, TSA and TPA are not statistically significant which disagrees with observations from EDA. The two only agree that NONS is a significant feature in determining VAT evasion. The results above also show that NLP, Tax office and Business Sector are also significant.

Further analysis was carried out to test the overall effect of Business Sector using Wald test. A chi-square test statistic of 30.2, with 5 degrees of freedom was associated with a p-value of 0.000013 indicating that the overall effect of business sector is statistically significant.

Therefore, Tax Office, Business sector, NONS and NLP were identified as significant features in the development of the models.

4.7.2 Model development

Logistic Regression model

Equation 3.1 was further applied to the training dataset with only the significant variables and the following results were obtained:

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.1092	-1.0714	-0.1994	1.169	1.5728

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.80384	0.22171	3.626	0.000288 ***
Tax.OfficeMTO	-0.43314	0.1134	-3.82	0.000134 ***
Business.SectorC	-0.62525	0.23483	-2.663	0.007756 **
Business.SectorF	-1.15594	0.25909	-4.462	0.000008137666 ***
Business.SectorG	-0.81304	0.22116	-3.676	0.000237 ***
Business.SectorH	-0.94586	0.27745	-3.409	0.000652***
Business.SectorO	-1.26472	0.27623	-4.579	0.00000468319 ***
NONS	0.13942	0.02205	6.324	0.000000000255 ***
NLP	0.05136	0.01747	2.94	0.003282 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2112.7 on 1523 degrees of freedom

Residual deviance: 2008.2 on 1515 degrees of freedom

AIC: 2026.2

Number of Fisher Scoring iterations: 4

Using Equation 2, the odds of the logit model features with outliers were obtained as below:

Table 4.4: Logistic Regression Odds Ratio

	OR	2.5%	97.5%
(Intercept)	2.2341047	1.4580967	3.4853141
Tax.OfficeMTO	0.6484676	0.5189369	0.8095415
Business.SectorC	0.5351273	0.3348593	0.8424319
Business.SectorF	0.3147626	0.1878603	0.5195883
Business.SectorG	0.4435095	0.2847471	0.6791588
Business.SectorH	0.3883463	0.2238545	0.6654026
Business.SectorO	0.2823196	0.1629177	0.4819203
NONS	1.1496049	1.1020755	1.2017832
NLP	1.0526984	1.0173707	1.0895408

From the logistic regression model above, holding other features constant, a taxpayer in MTO is less likely to be an evader compared to a taxpayer in LTO. From the tax experts' knowledge, this holds true since large taxpayers base on their large number of transactions to minimise their tax liability by over reporting their inputs. Further, holding other features constant; a unit increase in the number of no sales return filed increases odds of being a VAT evader by 14.96%, a unit increase in the number of late payments increases the odds of being a VAT evader by 5.27% while a VAT taxpayer in business sector A (Agriculture, Forestry, Fishing, Accommodation & Food Service Activities) is more likely to be a VAT evader compared to a taxpayer in any other business sector.

Further analysis was carried out to confirm if the developed logistic regression conforms to the assumptions of a binary logistic regression:

a. Binary Outcome. The class label has two levels that is "Evader" and "Non-Evader"
b. Lack of strongly influential outliers: Cook's Distance was calculated for all observations and 40 potential outliers were identified. Since no outlier had its cook's distance greater than 0.5, there are no influential values according to Harvey, Donato, Romme and Turner (2013).

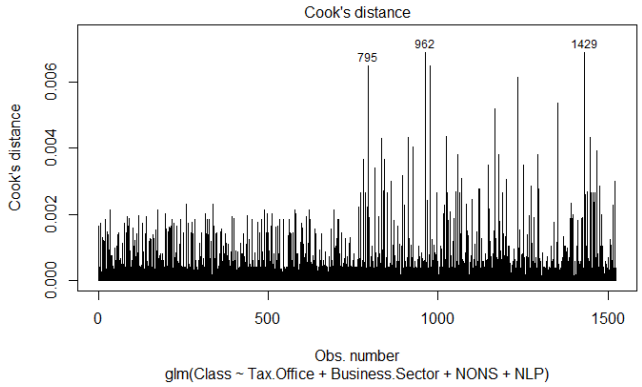
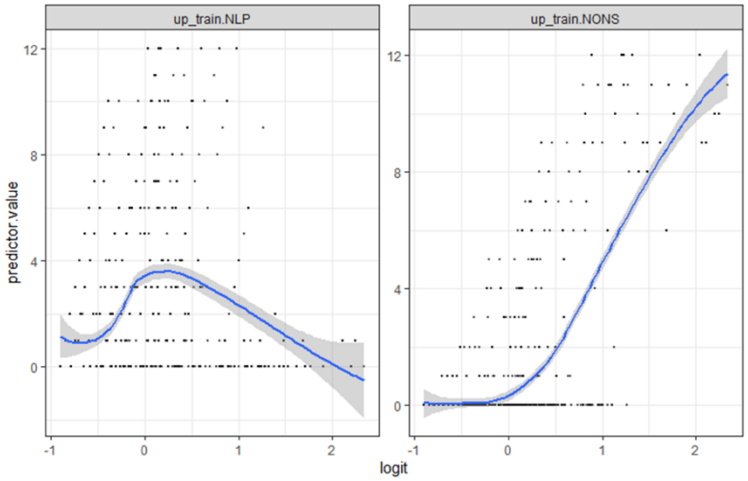


Figure 4.19: Cook's Distance plot

c. Linearity assumption of the logit to the continuous variables:



The smoothed scatter plots show that number of late payments (NLP) and number of no sales return (NONS) are both quite linearly associated with the VAT evader status outcome in logit scale. The plots highlight that as the number of late payments increase there are less chances of a taxpayer being a VAT evader and as the number of no sales returns increase the chances being a VAT evader increases. The number of no sales return plot agrees with taxation practice however, the late payment plot does not.

d. Absence of multi-collinearity

Multicollinearity occurs when two or more independent variables in a multiple regression model are correlated. Multicollinearity may result to the model giving invalid results about any individual attribute that is redundant with respect to others. Variance inflation factor (VIF) is a measure of

the amount of multicollinearity in multiple regression variables. Generalised Variance inflation factor (GVIF) for the significant features was calculated (Table 4.4)

Table 4.4: Variance Inflation Factor values

	GVIF	Df	GVIF^(1/(2*Df))
Tax Office	1.045456	1	1.022475
Business Sector	1.124924	5	1.011841
NONS	1.104444	1	1.050925
NLP	1.070276	1	1.034542

The Generalised-VIF values of all features are less than 5, which implies absence of multi-collinearity as recommended by Akinwande, Hussaini and Agboola (2015).

Finite Mixture of Logistic Regression Model

To select the most suitable number of components, the mixture is fitted an increasing number of components ranging from 2 to 4 and the Bayesian Information Criterion (BIC) is used to select the appropriate model (Grün & Leisch, 2007). The model with 2 components is selected since it has the lowest BIC (Grün & Leisch, 2007)

Table 4.5: FMLR with 2 to 4 components

	iter	converged	k	k0	logLik	AIC	BIC	ICL
2	128	TRUE	2	2	-142.243	322.4855	390.3564	456.1609
3	91	TRUE	3	3	-136.675	331.3507	434.9432	667.4147
4	200	FALSE	4	4	-135.748	349.4954	488.8094	850.0955

Ignoring the concomitant variable, the finite mixture of logistic regression model is fitted and below are the prior probabilities of the 2 components:

	prior	size	post>0	ratio
Comp.1	0.609	184	191	1
Comp.2	0.391	79	186	0.4

log Lik.' -142.2427 (df=19)
AIC: 322.4854 BIC: 390

Considering tax office as a concomitant variable, below are the prior probabilities:

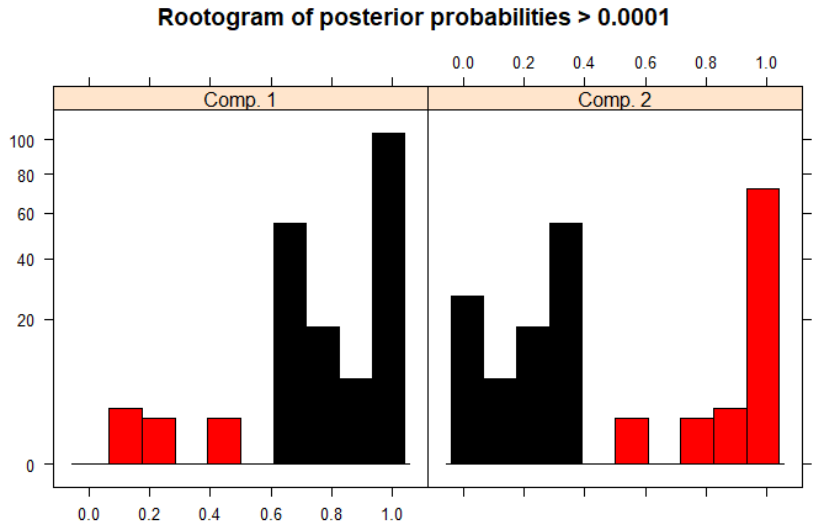
	prior	size	post>0	ratio
Comp.1	0.501	164	263	0.624
Comp.2	0.499	99	263	0.376

log Lik.' -156.7389 (df=17)
AIC: 347.4778 BIC: 408.2044

The full model without concomitant effect had a lower BIC and therefore preferred (Grün & Leisch, 2007).

Rootogram of the a-posteriori

A rootogram is a frequency graph where the axis is scaled by the square root of the frequencies. In finite mixture models, it can be used to see overlap between components. The rootogram of the a-posteriori probabilities below indicates that there is an overlap between component 1 and 2.



From Equation 4, the fitted parameters of the models are:

	Comp.1	Comp.2
coef.(Intercept)	-435.499859	55.478971
coef.Business.SectorC	4.060539	-45.514018
coef.Business.SectorF	-633.578742	-40.626138
coef.Business.SectorG	-664.052892	-43.100799
coef.Business.SectorH	419.175780	-46.885358
coef.Business.SectorO	-1151.445163	-55.203808
coef.Tax.OfficeMTO	433.975115	-14.075349
coef.NONS	317.646086	-4.188790
coef.NLP	-32.242434	1.751365

The above parameters indicate that holding all features constant, a VAT taxpayer in Component 1 is less likely to be an evader while one in Component 2 is more likely to be an evader. Holding other features constant, a taxpayer in MTO is more likely to be an evader compared to a taxpayer in LTO for Component 1 while it is otherwise for Component 2. A unit increase in the number of no sales return implies that a VAT taxpayer in Component 1 is more likely to be an evader however its otherwise for those in component 2. A unit increase in number of late payment implies that a VAT taxpayer in Component 1 is less likely to be an evader and its otherwise for those in Component 2. Holding other features constant, VAT taxpayers in Component 2 under Business Sector A (Agriculture, Forestry, Fishing, Accommodation & Food Service Activities) are more likely to evade VAT than taxpayers in other business sectors. Holding other features constant, with an exception of Business Sector C and H, a VAT taxpayer in Component 1 under business sector A is more likely to be a VAT evader than any other business sector. The agricultural and accommodation sector (Business Sector A) are majorly in informal sector with no proper transaction records and governing bodies. This makes them more vulnerable to tax evasion compared to other business sector which is line with the behavior in component 2. On the other hand, the manufacturing sector (Business Sector C) and transport sector (Business Sector H) transaction are complex and usually undermine the tax officers' ability to identify their transactions which agrees with Component 1 behavior.

4.7.3 Model performance comparison.

Test dataset containing 263 taxpayer information was used to compare the performance of LR and FMLR models. Comparison was based on measures obtained from the confusion matrix.

Table 4.5: Comparison of model classification performance

Classification Status	LR	FMLR
Correctly Classified	161	204
Incorrectly Classified	102	59
Total	**263**	**263**

Table 4.6: Comparison of model performance on testing dataset

Model	LR	FMLR
Accuracy	61.2%	77.6%
Precision	44.7%	69.6%
Recall	56.7%	61.1%

Table 4.7: Comparison of model performance on training dataset

Model	LR	FMLR
Accuracy	62.3%	78.9%
Precision	35.9%	59.9%
Recall	48.6%	68.5%

Table 4.6 highlights that consideration of 20% of the dataset as testing dataset evaluated on 3 performance measures resulted in FMLR having higher accuracy, precision and recall. Table 4.7 shows performance of the models on training dataset.

4.8 Discussion of Results

The study analyzed VAT evasion behavior of 1,304 taxpayers consisting of 376 evaders and 935 non-evaders each studied across 11 attributes. Using exploratory data analysis, returns not filed, number of late returns, total sales amount, average tax per month, number of no sales return were

identified as potential determinants of VAT evaders. Using backward elimination method with LR: tax office, business sector, number of no sales return and number of payments were identified as significant features in detection of VAT evaders. Since exploratory analysis is a data detection method that gives clues about the data (Tukey, 1977), inferential analysis significant results were considered.

Considering 20% of the dataset as testing dataset, FMLR with 77.6%, 69.6% and 61.1% obtained higher accuracy, precision and recall respectively than LR with 61.2%, 44.7% and 56.7% respectively. The results obtained after applying the models on training dataset (Table 12) imply that the two (2) models do not over-fit. These findings overall suggest that FMLR out performs LR.

The above performance is because logistic regression makes assumption of linearity between the dependent variable and the independent variables and no multi-collinearity between independent variables. Its better performance is determined by attainment of these assumptions which is nearly impossible in the field of taxation. On the other hand, Finite mixture models handle heterogeneity in the dataset than logistic regression.

CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 Introduction

This Chapter highlights the summary, conclusion and recommended future works from the study findings.

5.2 Summary

Tax authorities' objective is to maximize revenue collection in order to fund government expenditure. This goal is often affected by tax evaders who deliberately avoid paying their true tax liability through smuggling, submitting false tax returns or inaccurate financial statements, using fake documents to claim exemption and not reporting true income. Though, tax authorities have developed strategies to detect tax evaders, evasion cases have continuously grown. This research presents a comparison of LR and FMLR models in detecting VAT evaders in Uganda. A dataset of 1,311 VAT taxpayers, 448 from LTO and 863 from MTO. This dataset was partitioned into 80% training data and 20% testing dataset. The backward elimination method using LR was used for feature selection reducing the features from 11 to 4.

LR and FMLR models were developed to detect VAT evaders in Uganda. Though there is limited use of finite mixture models (FMLR) in detection of tax evaders as a data mining technique, they had higher accuracy (77.6%) than the LR model (61.2%). Other considered performance measures that is Precision and Recall support FMLR performance with 69.6% and 61.1% respectively verses 44.5% and 56.7% respectively for LR.

5.3 Conclusion

From the findings of the study, it can be concluded that Business sector, tax office, number of no sales returns and number of late payments are vital indicators in prediction of VAT evader. Among the developed models, FMLR detects VAT evaders with better accuracy, precision and recall than the LR. FMLR further had a better performance in the count of the correctly classified taxpayers than the LR.

5.4 Recommendations & Future study

Uganda Revenue Authority can utilize the findings of the study to detect VAT evaders for inclusion in the Audit plan with emphasis on the four (4) significant variables.

In regards to future research, other data mining techniques like random forest and artificial neural networks and features like assets and equity of a taxpayers may be applied to the same dataset and compare the results obtained. Researchers may explore other model evaluation metrics like ROC, AUC and time taken to develop each model to affirm model reliability in addition to the performance measures considered in this study.

CHAPTER SIX: APPENDICES

6.1 Data requisition letter

MAKERERE UNIVERSITY

Plot 51, Pool Road
P.O.Box 7062
Kampala, UGANDA
VOIP: 26105Cable MAKUNIKA

Phone: +256 414 541558/9
Fax: +256 414 530756
E-mail: hodsas@bams.mak.ac.ug
info@bams.mak.ac.ug
URL: www.bams.mak.ac.ug

College of Business and Management Sciences (CoBAMS)
School of Statistics and Planning
Department of Statistical Methods and Actuarial Science

September 23rd, 2020

The Assistant Commissioner Human Resource,
Uganda Revenue Authority.
P.O Box 7279 Kampala, Uganda.

Dear Sir/Madam,

Re: Request for Data for Research

Mr. Brian ApolloTusubira is a graduate student of Statistics (Registration Number 2018/HD06/1396U) in the Department of Statistical Methods and Actuarial Science, Makerere University.

He is requesting for data from your organisation, this will help him write his dissertation titled "Utilising Classification algorithms to detect VAT evasion in Uganda". He is obliged to adhere to all data requirements under your organisation and the Laws of Uganda. Any support rendered to him with respect to his research is highly appreciated.

Should you require further information, use kizomala@bams.mak.ac.ug or hodsas@bams.mak.ac.ug.

Yours sincerely

Saint Kizito Omala, Ph.D.
Chair and Lecturer, Department of Statistical Methods and Actuarial Science.

PS.: This letter doesn't bear the official stamp of the Department owing to the COVID-19 restrictions.

6.2 Acceptance letter

Head Office: Plot M193/M194 Nakawa Industrial Area
P.O.Box 7279, Kampala Uganda
Tel: +256417442097
Fax: +256414334419
Toll Free: 0800117000
Email: info@ura.go.ug

URA
Uganda Revenue Authority

URA/HRM/3.13.1

January 03, 2020

BRIAN APOLLO TUSUBIRA
C/O Makerere University
P O Box 7062,
KAMPALA, UGANDA

Dear Sir,

LETTER OF OFFER

Please refer to your request to carry out research on the topic, "**To Develop a Supervised Machine Learning Model that Predicts Value Added Tax (VAT) Evaders in Uganda: A case Study of Uganda Revenue Authority**".

This is to inform you that your request has been granted on the following terms:

a) Your research period shall not exceed two months. If you require more time, then you shall formally request the Assistant Commissioner Human Resources.
b) You will also avail a copy of research results in a bound book to the Manager Human Resource Development after completion of research.
c) You will sign an oath of secrecy to maintain confidentiality of information received in the course of the research.

Your research will be guided by the heads of station where you will issue questionnares,carry out interviews and you are obliged to agree on how the research will be conducted.

I wish you success in your endeavours.

Yours faithfully

Alice Doreen Nansamba
Ag.MANAGER HUMAN RESOURCE DEVELOPMENT

REFERENCES

Akinwande, M. O., Hussaini, D. G., & Agboola, S. (2015, December). Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. *Journal of Statistics, 5*, 754-767. doi:doi.org/10.4236/ojs.2015.57075

Ali, H., Salleh, M. N., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019, June). Imbalance class problems in data mining: a review. *Indonesian Journal of Electrical Engineering and Computer Science, 14*(3), 1560-1571. doi:10.11591/ijeecs.v14.i3.pp1560-1571

Almunia, M., Gerard, F., Hjort, J., Knebelmann, J., Nakyambadde, D., Raisaro, C., & Tian, L. (2017). *An analysis of discrepancies in tax declarations submitted under value-added tax in Uganda, S-43312-UGA-1.* Kampala: International Growth Centre.

Assylbekov, Z., Melnykov, I., & Bekish, R. (2016). Detecting Value-Added Tax Evasion by Business Entities of Kazakhstan. *International Conference on Intelligent Decision Technologies* (pp. 37-49). Springer, Cham.

Cobham, A., & Janský, P. (2017). Measuring misalignment: The location of US multinationals' economic activity versus the location of their profits 39(1). *Development Policy Review*, 91-110.

Crivelli, E., Mooij, R. D., & Keen, M. (2015). *Base Erosion, Profit Shifting and Developing Countries.* Washington, DC: International Monetary Fund.

Egger, P. H., Merlo, V., & Wamser, G. (2018). Unobserved tax avoidance and the tax elasticity of FDI. *Journal of Economic Behavior & Organization*, 1-18.

González, P. C., & Velásquez, J. D. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications, 40 (5)*, 1427–1436.

Grün, B., & Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. Computational Statistics & Data Analysis, 51(11). *Computational Statistics & Data Analysis, 51*(11), 5247–5252. doi:doi:10.1016/j.csda.2006.08.014

Ha, R., Chang, P., Karcich, J., Mutasa, S., Sant, E. P., Liu, M. Z., & Jambawalikar, S. (2018). Convolutional Neural Network Based Breast Cancer Risk Stratification Using a Mammographic Dataset. *Academic Radiology*, 1-6.

Harvey, B. J., Donato, D. C., Romme, W. H., & Turner, M. G. (2013). Influence of recent bark beetle outbreak on fire severity and postfire tree regeneration in montane Douglas-fir forests. *Ecological Society of America (ESA), 94*(11), 2475–2486. doi:https://doi.org/10.1890/13-0188.1

Hutton, E., Thackray, M., & Wingender, P. (2014). *Revenue Administration Gap Analysis Program-The Value-Added tax Gap.* Kampala, Uganda: International Monetary Fund (IMF).

Jihal, H., Talhaou, M. A., Daif, A., & Azzouazi, M. (2018). Predictive Analytics as A Service on Moroccan Tax Evasion. *International Journal of Engineering & Technology, 7 (4.32)*, 90-92.

Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200-1205. doi:doi:10.1109/mipro.2015.7160458

Jupri, M., & Sarno, R. (2018). Taxpayer compliance classification using C4.5, SVM, KNN, Naive Bayes and MLP. *International Conference on Information and Communications Technology (ICOIACT)* (pp. 297-303). Yogyakarta, Indonesia: IEEE.

Lahann, J., Scheid, M., & Fettke, P. (2019). Utilizing Machine Learning Techniques to Reveal VAT Compliance Violations in Accounting Data. *2019 IEEE 21st Conference on Business Informatics (CBI)* (pp. 1 - 10). Moscow, Russia: IEEE.

Luigidell'Olio, AngelIbeas, Juan deOña, & Rocio deOña. (2018). Public Transportation Quality of Service. *Elsevier*, 101-139.

Mathews, J., Mehta, P., Kuchibhotla, S., Bisht, D., Chintapalli, S. B., & Rao, S. V. (2018). Regression Analysis towards Estimating Tax Evasion in Goods and Services Tax. *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 758-761). Santiago, Chile: IEEE.

Mawejje, J., & Okumu, I. M. (2016). Tax Evasion and the Business Environment in Uganda. *South African Journal of Economics*, 341-498.

McLachlan, G., & Peel, D. (2000). Finite Mixture Models. *John Wiley & Sons Hoboken*.

Mehta, P., Mathews, J., Suryamukhi, K., & San, K. (2018). Predictive Modeling for Identifying Return Defaulters in Goods and Services Tax. *International Conference on Data Science and Advanced Analytics* (pp. 631-637). Turin, Italy: IEEE.

Qasim, O. S., & Algamal, Z. Y. (2018). Feature selection using particle swarm optimization-based logistic regression model. *Chemometrics and Intelligent Laboratory Systems, 182*, 41-46. doi:https://doi.org/10.1016/j.chemolab.2018.08.016

Ramya, R. S., & Kumaresan, S. (2015). Analysis of feature selection techniques in credit risk assessment. *2015 International Conference on Advanced Computing and Communication Systems* (p. doi: 10.1109/ICACCS.2015.7324139). IEEE.

Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017, Jul-Sep). Common pitfalls in statistical analysis: Logistic regression. *Perspect Clin Res, 8*(3), 148–151. doi:doi: 10.4103/picr.PICR_87_17

Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques, 50. *Decision Support Systems*, 491-500.

Roux, D. d., Perez, B., Moreno, A., Villamil, M. D., & Figueroa, C. (2018). Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach. *KDD '18 Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 215-222). London, United Kingdom: ACM New York, NY, USA ©2018.

Sahin, Y., & Duman, E. (2011). Detecting credit card fraud by ANN and logistic regression. *International Symposium on Innovations in Intelligent Systems and Applications.*, 315-319. doi:doi:10.1109/inista.2011.5946108

Shadi, A., Monther, A., & Muneer, Y. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. *Journal of Computational Science*, 152-160.

Sharma, A., & Panigrahi, P. K. (2012). A Review of Financial Accounting Fraud Detection based on Data Mining Techniques. *International Journal of Computer Applications, 39(1)*, 37-47.

Tang, J., Alelyani , S., & Liu, H. (2014). *Feature Selection for classification: A Review.* Dehli: College of Vocational Studies.

Tukey, J. W. (1977). *Exploratory data analysis. Reading, MA: Addison Wesley.*

Uganda Revenue Authority [URA]. (2011). *Taxation handbook: a guide to taxation in Uganda.* Kampala: Fountain Publishers.

Uganda Revenue Authority [URA]. (2013). *Tax to GDP Ratio: A comparative study of Uganda with selected East Africancountries and South Africa.* Kampala: Uganda Revenue Authority.

Uganda Revenue Authority [URA]. (2018). *Revenue Performance report FY 2017/18.* Kampala, Uganda: Uganda Revenue Authority.

Unler, A., & Murat, A. (2010). A discrete particle swarm optimization method for feature selection in binary classification problems. *European Journal of Operational Research, 206*(3), 528–539. doi:doi:10.1016/j.ejor.2010.02.032

Verma, I. S. (2015). Expansion, Impact and Challenges of IT & CS: Knowledge Data Discovery and Its Issues. *10th Biyani International Conference (BICON-15). 1*, pp. 88-91. Jaipur (India): Biyani Institute of Commerce & Management Pvt. Ltd.

Wu, R.-S., Ou, C., Lin, H.-y., & Chang, S.-I. (2012). Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications, 39(10)* (pp. 8769–8777). Elsevier Ltd.