Enhancing Taxpayer Compliance through the Integration of Predictive Modelling and Digital Insights

Opeyemi Olaosebikan

University of Wolverhampton

Faculty of Science and Engineering

MSC Thesis 2024

Declaration

Enhancing Taxpayer Compliance through the Integration of Predictive Modelling and Digital Insights

Opeyemi Olaosebikan

University of Wolverhampton

Faculty of Science and Engineering

A thesis submitted in partial fulfilment of the requirements of the University of Wolverhampton for the Masters Degree of Data Science (MSC)

JUNE 2024

This work and any part thereof have not previously been presented in any form to the university or to any other body whether for assessment, publication or for any other purpose (unless otherwise indicated). Save for any express acknowledgments, references and/or bibliographies cited in the work, I can confirm that the intellectual content of this work is the result of my own efforts, not those of any other person.The right of Opeyemi Olaosebikan to be identified as the author of this work is asserted in accordance with ss. 77 and 78 of the Copyright, Design and Patent Act 1988. At this date, the author owns the copyright.

Signature: Opeyemi Olaosebikan

Date: July 12th, 2024

Abstract

Effective tax compliance strategies are crucial for governments to ensure sustainable revenue generation and foster taxpayer engagement. This study aims to enhance taxpayer compliance in Nigeria by integrating advanced predictive modelling techniques with digital insights and demographic indicators. Leveraging historical tax records and demographic data from the Federal Inland Revenue Service (FIRS), the research employed the XGBoost, Random Forest and Decision Tree classifiers to develop robust predictive models. The impact of digitalization on tax compliance was a focal point, examining electronic filing, online payments and engagement with platforms like Taxpromax and Remita.

The findings demonstrated the strong performance of the XGBoost model in an imbalanced dataset, with the default and hyperparameter-tuned versions achieving high accuracy (84.86%), precision (91.40%), recall (67.61%), and ROC-AUC (92.94%) scores. The models' superior ability to detect non-compliant or late compliance cases highlights its potential value for tax authorities in identifying high-risk taxpayers despite a nuanced trade-off with recall/precision scores.

The study also provides valuable insights into the factors influencing tax compliance, aligning with existing literature. It revealed that the integration of digital insights from electronic filing and online payment patterns significantly influences compliance behaviour, with higher engagement on digital platforms correlating with improved compliance rates.

Based on these findings, the study recommends that tax authorities like the FIRS adopt data-driven strategies, leverage digital resources effectively and implement targeted compliance initiatives tailored to specific taxpayer segments. This approach can enable the FIRS to enhance taxpayer engagement, improve compliance, and optimize revenue collection, aligning with its strategic goals.

The expected outcomes include improved predictive precision, a comprehensive understanding of the digital transformation's impact on taxpayer behaviour in Nigeria, and the development of customized enforcement approaches tailored to specific taxpayer segments. The research findings have the potential to enable Nigerian tax authorities to adopt data-driven strategies, leverage digital resources effectively, and implement targeted compliance initiatives, ultimately fostering a more efficient and equitable tax ecosystem, as emphasized by the African Tax Administration Forum's initiatives on tax compliance. (ATAF, 2020).

Table of Contents

Table of Figures

Index of Tables

Chapter 1: Introduction

The introductory chapter of this study covers the background and justification for the research, establishing its relevance and necessity. It outlines the research aim, objectives, and questions that guide the investigation. The chapter details the adopted methodology, explaining the techniques and approaches used, and reviews the contributions to existing knowledge, emphasizing the study's significance. Additionally, it discusses the scope and limitations of the study, acknowledging potential constraints. Finally, the chapter presents the overall structure of the thesis, providing a roadmap for the reader.

Research Background

Tax compliance behaviour is believed to be influenced by the government's provision of basic infrastructure. However, Nichita and Batrancea (2012) also identified various determinants of tax compliance behavior, including social, psychological, political, industrial, business, and economic factors.

Efunboade (2014) noted that initiatives such as tax week, tax counseling and education, incentives for early tax return filing, penalty provisions, regular audits, and the development of electronic tax management systems can enhance revenue satisfaction in developing countries through self-assessment. Similarly, Okello (2014) pointed out that education, a service-oriented approach, stringent deterrents to non-compliance, regular audits, and transparency from the government and tax authorities can boost voluntary compliance and improve revenue generation.

Effective tax administration is crucial for governments to ensure sustainable revenue generation and adequate funding for public services and infrastructure. In Nigeria, tax revenues contribute significantly to the country's fiscal resources, with the Federal Inland Revenue Service (FIRS) serving as the principal agency responsible for tax collection and compliance.

Historically, Nigeria's tax system has faced numerous challenges, including low compliance rates, inadequate enforcement mechanisms and a lack of comprehensive data and analytics capabilities (ATAF, 2019). These challenges have hampered the government's ability to optimise revenue collection and effectively allocate resources for national development.

In recent years, the Nigerian government has recognized the importance of leveraging technology and data-driven approaches to enhance tax compliance and taxpayer engagement. The advent of digital platforms, online filing systems, and electronic payment methods has opened up new avenues for streamlining tax administration processes and improving transparency (FIRS, 2021).

The increasing availability of diverse data sources, such as historical tax records, demographic information and economic indicators has also created opportunities for developing advanced predictive models to forecast taxpayer behaviour and compliance trends (Ullah et al., 2021). By integrating machine learning techniques with these data sources, tax authorities can obtain valuable insights and develop targeted strategies to enhance compliance and taxpayer engagement.

Additionally, the impact of socioeconomic factors on taxpayer behaviour has been widely recognized (IMF, 2021) with variables such as employment rates, income distribution, and

consumer spending patterns having significant influence on compliance tendencies across different demographic groups and economic contexts. Incorporating these factors into predictive models can enhance their accuracy and provide a more nuanced understanding of taxpayer behaviour.

Research Justification

Nigeria, a West African nation situated along the Gulf of Guinea, is a federal republic consisting of 36 states and the Federal Capital Territory, Abuja. English serves as its official language, and its currency is the Nigerian naira (NGN). As of early 2024, Nigeria's population reached 226.5 million, according to DataReportal. While it rebased its GDP in 2014, becoming Africa's largest economy at the time, its economic landscape has since evolved. Despite a drop in its global ranking Nigeria remains a significant oil producer in Africa and a member of OPEC, recognized for its growth potential and facing challenges such as fluctuating oil prices, inflation and low technology adoption as reflected in the interest rate trends reported by Trading Economics.

Despite these hurdles, taxation of the non-oil sector remains a pressing issue, with low compliance rates attributed to weak revenue administration, tax evasion and a lack of data, particularly in the informal sector. To address these concerns, the Presidential Fiscal Policy and Tax Reforms Committee was inaugurated in August 2023, focusing on fiscal governance, revenue transformation, and economic growth facilitation. The impact of the COVID-19 pandemic on Nigeria's economy and tax revenues is also a key consideration in understanding the country's current economic landscape.

Tax compliance plays a pivotal role in ensuring the financial well-being of nations, as governments rely heavily on tax revenues to fund public goods and services. In Nigeria, the Federal Inland Revenue Service (FIRS) oversees the collection of various taxes, including Pay As You Earn (PAYE), Stamp Duty, Companies Income Tax, Value Added Tax (VAT), Personal Income Tax, and Petroleum Profit Tax (FIRS, 2022). Accurate forecasting of tax revenues is crucial for effective budgeting, resource allocation, and aligning fiscal policies with the country's economic landscape (International Monetary Fund, 2018). Traditionally, classical statistical models such as Seasonal Autoregressive Integrated Moving Average (SARIMA) and Autoregressive Integrated Moving Average (ARIMA) have been employed for tax revenue forecasting. However, with the advent of machine learning and the availability of diverse data sources, there is a growing opportunity to leverage advanced predictive modelling techniques and integrate multidimensional data to enhance the accuracy and insights of tax compliance forecasts (Jang, 2019; Ullah et al., 2021).

This study aims to develop a comprehensive approach to predicting taxpayer compliance in Nigeria by integrating predictive modelling techniques with digital insights and socio-demogrpahic indicators. By harnessing the power of machine learning algorithms and leveraging historical tax data, demographic information, economic indicators, and digital interaction statistics, the research seeks to create robust predictive models capable of anticipating taxpayer behaviour and compliance trends.

The integration of digital knowledge is a key aspect of this study, as it investigates how the proliferation of digital platforms, electronic filing systems, and online payment methods influence taxpayer engagement and compliance. By analysing patterns in digital interactions,

the research aims to uncover valuable insights that can inform targeted engagement strategies and enhance the overall tax ecosystem (OECD, 2020; ATAF, 2021).

Ultimately, this study seeks to contribute to the field of tax compliance by offering a comprehensive framework that integrates predictive modelling, digital insights, and demographic information. The findings have the potential to empower Nigerian tax authorities with data-driven strategies, enabling them to optimise revenue collection, foster taxpayer engagement, and promote a more efficient and equitable tax system.

Research Aims and Objectives

The project aims to improve the precision and efficiency of predicting tax compliance by combining machine learning techniques with information from digitalization and sociodemographic indicators. By using past tax data, demographic details, economic indicators, and digital interaction statistics, the goal is to create strong predictive models that can forecast taxpayer actions. Through the utilisation of digital resources and examining how they influence compliance; the objective of this study is to offer practical approaches for tax agencies to boost revenue collection and improve taxpayer involvement.

The specific objectives of this study are as follows:

- Develop Advanced Predictive Models: Employ the use of sophisticated machine learning algorithms to build predictive models that can anticipate how taxpayers will comply with tax regulations. These models will be based on complex datasets that include historical tax records and demographic information.
- Integrate Digital Insights: By analysing data from digital platforms, trends in electronic

filing, and patterns of online payments we examine how digitalization affects taxpayer behaviour and compliance. This information is then used to improve the accuracy of predictive models and identify opportunities for proactive engagement.

- Improve Engagement Strategies: Leverage predictive insights to develop effective strategies for tax authorities to engage with taxpayers and improve compliance. Recommend customised approaches for communication, outreach, and enforcement based on the results of predictive modelling and findings from digital integration

Key Questions Answered by Research on Tax Compliance Prediction

Given the research background, justification, aim and objectives, the research provided answers to the following questions:

How can advanced predictive models improve tax compliance in Nigeria?

The research demonstrated that advanced machine learning models like XGBoost can significantly enhance the ability to predict taxpayer compliance, identifying non-compliant taxpayers with high accuracy and precision.

What is the impact of digital engagement on taxpayer compliance?

The study found that higher engagement with digital platforms such as the FIRS e-Services portal and online payment systems is correlated with improved compliance rates, highlighting the importance of digital transformation in tax administration.

Which predictive model is more effective for identifying non-compliant taxpayers?

The research compared the default XGBoost model and a tuned XGBoost model. It was determined that the default model, with its higher accuracy and recall, is more effective for identifying non-compliant taxpayers.

What are the key metrics to evaluate the performance of predictive models in tax compliance?

The study used metrics such as accuracy, precision, recall, F1-score, F-beta score, and ROC-AUC to evaluate model performance, providing a comprehensive assessment of each model's strengths and weaknesses.

How do demographic factors influence tax compliance behaviour?

The integration of demographic information in predictive models highlighted that certain demographic segments are more likely to be non-compliant, allowing for more targeted compliance strategies.

What recommendations can be made to tax authorities for improving compliance?

The research provided several practical recommendations, including the use of the default XGBoost model, enhancing digital engagement, implementing targeted campaigns, continuous monitoring of predictive models, and developing customized communication strategies.

How does the predictive precision of the models compare to previous studies on tax compliance?

The research compared its findings with previous studies, demonstrating that the integration of advanced machine learning algorithms and digital insights can significantly improve predictive precision and overall compliance rates.

What role does electronic filing and online payment play in tax compliance?

The study confirmed that patterns in electronic filing and online payments are significant predictors of tax compliance, emphasizing the need for tax authorities to promote and facilitate these digital methods.

By addressing these questions, the research provides a detailed understanding of how predictive modeling and digital integration can be leveraged to enhance tax compliance, offering actionable insights for tax authorities in Nigeria and potentially other regions with similar challenges.

Significance and Potential Contributions of the Study.

This study carries substantial significance and has the potential to make valuable contributions towards improving/forecasting taxpayer engagement and compliance in Nigeria. Through the incorporation of predictive modelling techniques, digital insights, and socio demographic factors, the study endeavours to establish a comprehensive framework that has the capability to transform the approach of tax authorities towards compliance strategies.

Firstly, developing predictive models using advanced machine learning algorithms and gradient boosting techniques has made it possible to accurately predict taxpayer compliance behaviour. This is a significant tool that has the potential to help tax authorities to proactively anticipate compliance trends and take timely actions to enforce tax laws on specific groups of taxpayers.

Secondly, the incorporation of digital insights plays a crucial role in gaining a deeper understanding of how the proliferation of digital platforms and electronic services impacts taxpayer engagement and compliance. Through the analysis of patterns in electronic filing, online payments and digital interactions, this study offers valuable insights into the effects of digital transformation on the tax ecosystem. These findings can guide tax authorities in optimising their digital strategies, improving user experiences and utilising technology to promote greater compliance and simplifying tax processes for taxpayers.

Additionally, the study's findings and recommendations hold relevance for policy decisions and strategic planning by tax authorities in Nigeria. By offering evidence-based insights and actionable recommendations, this research can contribute to the development of effective tax policies, legislative reforms and the implementation of best practices aligned with international standards and initiatives, such as those advocated by the African Tax Administration Forum (ATAF) and the Organisation for Economic Co-operation and Development (OECD).

Ultimately, this study aims to improve the tax system in Nigeria by increasing efficiency, transparency, and fairness. It also aims to encourage taxpayers to participate, follow tax laws willingly, and help the economy grow sustainably. The findings of this study can be applied to

different fields beyond taxes, such as predictive modelling, digital transformation, and socioeconomic influences on human behaviour and organisational plans.

Thesis Overview

Figure 1: Thesis Overview

This report is systematically structured to provide a logical understanding of how the research objectives were achieved. The thesis begins with an introductory section in Chapter 1 and concludes with the conclusions and recommendations in Chapter 6.

Figure 1 provides a visual representation of the thesis structure, showing its organization.

A detailed explanation of the thesis structure follows:

Chapter 1 discusses the background and justification of the study, presenting the research aim, objectives, questions, and methodology. It also covers the study's contribution to knowledge, its scope and limitations, and an outline of each chapter.

Chapter 2 reviews relevant literature and theoretical framework on tax complaince. Its also discusses key concepts, methodologies and findings and limitations of previous studies on tax compliance as it relates to the Nigerian economy.

Chapter 3 discusses the methodological approach used to achieve the study's aim and objectives. The chapter discusses the dataset used in this research, it gives an explanation on data collection, its sources and the ethical consideration adopted. It also discusses the dat analysis techniques implored and how it addresses the research objectives.

Chapter 4.Data Analysis and Results: This chapter outlines the meticulous data pre-processing steps undertaken, including data cleaning and transformation, to ensure the integrity and validity of the analysis. It then describes the analytical process in detail, providing a clear narrative of the techniques and tools employed.

The results are presented with clarity, supported by appropriate visualizations to enhance the comprehension of the findings. Each result is carefully interpreted in light of the research questions posed at the beginning of the study.

The chapter proceeds to discuss the significance of these results, particularly in the context of the research objectives, highlighting how the findings contribute to the broader field of study. Additionally, it addresses any unexpected results, offering thoughtful explanations that may account for these anomalies.

This summary encapsulates the critical elements of the data analysis and results chapter, providing a snapshot of the methods and outcomes of the research.

Chapter 5 Discussion and Interpretation: This section provides a critical interpretation of the research findings, contextualized within the broader literature reviewed earlier in the thesis. It candidly addresses the limitations and potential biases inherent in the chosen methodology and data set. The results are then juxtaposed with existing studies, drawing attention to any novel insights or contributions that this research offers. Finally, the chapter explores the practical implications of the findings, considering their relevance and potential impact on real-world applications. This discussion not only reinforces the validity of the research but also propels it into the realm of practical utility.

Chapter 6 The concluding chapter synthesizes the key findings of the research, underscoring their contributions to the evolving landscape of data science and artificial intelligence. It reaffirms the significance of the study, emphasizing its potential to influence future developments and applications within the field. The chapter concludes with a forward-looking perspective, suggesting avenues for subsequent research that build upon the insights garnered from this work. This final section encapsulates the essence of the research, its value to the academic community, and its prospective utility in practical scenarios.

Chapter 7 References Presents comprehensive list of all references cited in this dissertation.

Using Harvard Referencing Style.

Chapter 2: Literature Review

This chapter discusses crucial concepts, theories, methodologies and findings from previous studies, underscoring their significance and applicability to the current study. It presents an extensive analysis of relevant literature and theoretical frameworks concerning tax compliance, predictive modelling, digital transformation, and sociodemographic factors.

Introduction

Tax revenue is fundamental to the sustainability of any economy. According to the Organization for Economic Co-operation and Development (OECD) in 2018, the top ten global economies have an average tax-to-GDP ratio of 33.8%. In contrast, the average tax-to-GDP ratio for the 26 largest economies in Africa during the same period is 17.2%, with Nigeria significantly lower at approximately 6%. This disparity is a contributing factor to Nigeria's increasing debt levels and, if not addressed, poses a threat to the country's economic sustainability (Audu, 2023).

Tax compliance refers to the voluntary submission of all necessary tax documents within the stipulated timeframe by taxpayers, along with accurate declaration of their tax obligations following the relevant legal principles, guidelines, and judicial rulings (Roth et al., 1989; Imas and Mareska).

Tax compliance encompasses the timely submission and reporting of tax returns, including the overall number of successfully submitted tax documents and the proportion of tax return forms filed punctually (Tilahum, 2018)

Tax compliance involves taxpayers and tax compliance intermediaries adhering to their responsibilities (Mohammed et al., 2016; James and Alley, 2004).

On the other hand, tax non-compliance encompasses various forms of deviant behaviour, such as Failure to file tax returns as Taxpayers may not file their tax returns within the required timeframe, leading to late compliance or non-compliance (Kirchler et al., 2008) also Non-payment of taxes by taxpayers who fail to pay the full amount of taxes owed, either intentionally or due to financial constraints (Alm, 2019).

The legal and institutional framework for regulating tax avoidance and evasion in Nigeria includes tax laws, regulations and enforcement mechanisms by tax authorities such as the Federal Inland Revenue Service (FIRS) (Ibrahim et al., 2014; Modugu and Anyaduba, 2014).

Tax compliance and revenue generation remain critical challenges for the Nigerian government, as highlighted in the Finance Act 2021. Ensuring effective tax compliance is essential for funding public services, infrastructure development, and economic growth.

However, traditional approaches to tax administration and enforcement have faced limitations, necessitating the exploration of more innovative and data-driven strategies. This research aims to develop an integrated approach that combines predictive modelling techniques, digital insights and sociodemographic factors to enhance tax compliance strategies and revenue collection in Nigeria.

The potential contributions of this research are multifaceted. Theoretically, it seeks to advance the understanding of taxpayer behaviour by integrating insights from diverse Lugbe MSTO data sources and analytical techniques. Practically, the development of advanced predictive models can enable tax authorities forecast revenue more effectively, target high-risk groups and tailor engagement strategies to specific taxpayer segments. Furthermore, by leveraging digital

insights and demographic factors, this research can inform the design of user-friendly digital platforms and policy interventions that address the unique challenges faced by different demographic groups.

Theoretical Framework

Taxpayer non-compliance, whether in the form of underreporting, late filing, or non-payment of taxes, can lead to significant revenue losses and undermine the fairness and effectiveness of the tax system (Alm, 2019; Kirchler, 2007).Tax authorities are now increasingly adopting the integration of predictive modelling and digital insights and historical records to enhance their compliance strategies.

The study of tax compliance behaviour has been informed by various theoretical frameworks, each offering valuable insights into the factors that influence taxpayer decisions and actions.

- The theory of planned behaviour (Ajzen, 1991) suggests that individuals' intentions and behaviours are shaped by their attitudes, subjective norms and perceived behavioural control. This theory has been applied to understand tax compliance, where attitudes towards taxation, social norms, and perceived ability to comply influences taxpayers' intentions and subsequently their actions (Bobek et al., 2007).
- Closely related is the conventional tax compliance theory which has its origins in the work of Adam Smith in 1776. This theory suggests that taxpayers evaluate the benefits of evading taxes against the risks of detection and punishment by tax authorities (Kirchler, 2007). Taxpayers are more likely to comply with tax laws when they perceive

the costs of non-compliance, such as penalties and the risk of detection, to be higher than the benefits of evasion.

- Another relevant theoretical perspective is the slippery slope framework (Kirchler et al., 2008), which posits that tax compliance is influenced by both enforced and voluntary motivations. Enforced compliance is driven by the perception of authorities' power and the potential for deterrence through audits and penalties. In contrast, voluntary compliance is fostered by trust in authorities and the perceived legitimacy of the tax system. The framework suggests combination of both coercive and cooperative strategies in promoting tax compliance.

These theories suggest that individuals may engage in tax evasion when they perceive the benefits of non-compliance to outweigh the risks, they also highlight the effectiveness of deterrence mechanisms such as tax audits and penalties as crucial in discouraging tax avoidance and promoting voluntary compliance. While these theories offer valuable insights by highlighting the importance of audits and penalties and the likelihood of detection in shaping taxpayer behaviour, they may not fully capture the complexities of taxpayer behaviour in the Nigerian context, where factors such as informal economic activities, low tax morale, and limited digital infrastructure play significant roles. This research seeks to integrate existing theoretical frameworks with contextual factors and emerging data sources to develop a more comprehensive understanding of tax compliance behaviour in Nigeria.

Overview of Previous Research

The existing body of research on tax revenue forecasting and tax compliance prediction has provided valuable insights and methodological approaches that inform the current study on "Enhancing Taxpayer Compliance through the Integration of Predictive Modelling and Digital Insights."

Previous studies have explored various factors influencing tax compliance behavior. Kirchler (2007) and Batrancea et al. (2019) emphasized the importance of taxpayer attitudes, perception of fairness, and the effectiveness of enforcement mechanisms in shaping compliance decisions. Alm and Torgler (2012) highlighted the role of ethics and the fear of detection and punishment as key determinants of early compliance. Researchers have also employed a range of predictive modelling techniques to forecast tax compliance and tax revenue. Olsen et al. (2016) utilized logistic regression and decision tree models to predict tax evasion, while Duan et al. (2020) explored the application of machine learning algorithms, such as random forests and gradient boosting, for tax compliance prediction.

Building on this foundation, the following section provides a more detailed examination of the key studies that have informed the current investigation.

The study conducted by Abdu Masanawa Sagir et al. aimed to forecast Nigerian economic growth based on corporate income tax (CIT) data using artificial neural networks (ANNs) and multiple linear regression (MLR) models. The researchers employed three different ANN training algorithms: conjugate gradient back-propagation with Fletcher-Reeves restarts, Bayesian regularization, and gradient descent with an adaptive learning rate. The input variables used in the models were agricultural & plantation (AP), banks & financial institutions

(BFI), breweries, bottling & beverage (BBB), hotels & catering (HC), and professional service telecommunication (PST), while the output variable was the CIT collected in Nigeria from 2015 to 2020.

The detailed results showed that:

- For the ANN models: Conjugate gradient back-propagation had an overall regression of 0.97529 and a mean squared error (MSE) of 1.7486e+21. Bayesian regularization had the best performance with an overall regression of 0.99999 and an MSE of 1.3007e+12 and Gradient descent with adaptive learning rate had an overall regression of 0.99355 and an MSE of 9.0901e+19.
- For the MLR model: The model had an R-squared of 90.81% and an adjusted R-squared of 82.24%, indicating a good fit.

The results suggest that while both ANN and multiple linear regression can be effective in forecasting economic growth based on CIT data, the Bayesian regularization approach stands out as the most robust and accurate method among the tested ANN algorithms. The authors recommend future research to focus on validating these models with larger datasets and exploring additional variables to enhance the predictive power and generalizability of the models.

Another study by Mohammad Zoynul Abedin et al. focused on predicting tax defaults using feature transformation-based machine learning methods. The research utilized a dataset on Finnish limited liability firms, revealing that approximately 12% of active Finnish enterprises had

outstanding taxes at the end of 2015, totalling over three billion Euros in overdue corporate taxes. The study also references World Bank statistics indicating that around 40% of firms globally pay their taxes, whereas 60% fail to comply, with these unpaid taxes potentially remaining unrecovered in future years. Furthermore, the study underscores the global rise in tax default rates.

The authors proposed an automated tax default prediction system that incorporates advanced data analytic techniques with financial predictors derived from corporate financial statements. Various feature transformation techniques, including log transformation, Z-normalization, scaled transformation, sine transformation, and square-root transformation, were applied to the dataset to enhance the informational value of the financial indicators. These transformed datasets were then rigorously tested using 13 different machine learning algorithms to detect tax defaults both in the default year and one year prior.

The results of the study show that the log-transformation method ranked first for both accuracy and F-measure in predicting tax defaults in the default year and one year prior. The square-root transformation ranked second in terms of performance.

Furthermore, the study demonstrates that ensemble learning methods, such as XGBoost, Gradient Boosting (GB), Systematic Forest (SysFor), and Forest PA (ForestPA), outperformed the single classifiers, except for Logistic Regression (LR), which performed well for the default year prediction. XGBoost was identified as the best-performing model in terms of accuracy and F-measure for predicting tax defaults in the default year, while SysFor dominated for the one-year-ahead tax default prediction.The random forest (RF) model was also used to verify the

importance of financial indicators for tax default prediction, and the study found that the equity ratio, liquidity ratios (quick ratio and current ratio), and debt-to-sales ratio were the most important indicators of tax defaults.Additionally, the study shows that XGBoost and GB achieved the lowest expected misclassification cost (EMCC), providing the best trade-off between false-positive and false-negative errors under different misclassification costs. XGBoost was particularly effective in predicting tax defaults in the real-world scenario of varying misclassification costs, leading to substantial cost reduction for decision-makers.

In summary, the study underscores the critical importance of well-designed tax default prediction systems that incorporate feature transformation and advanced machine learning methods. Implementing an automated tax default prediction system effectively has significant implications for tax administration, as it can improve the accuracy and efficiency of tax compliance management.

Hamza Erdoğdu and Recep Yorulmaz also compared the performance of three different time series forecasting models - Random Walk, SARIMA, and BATS - in predicting monthly tax revenues in Turkey. The researchers used a dataset that was split into a training set covering the period from 2006:01 to 2014:12 and a testing set covering the period from 2015:01 to 2018:12.The study evaluated the forecasting accuracy of the models using several performance metrics, including Mean Error (ME), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), and Theil's U. The results showed that the BATS model outperformed the

Random Walk and SARIMA models in terms of all seven evaluation criteria.Specifically, the BATS model had the lowest ME (2,042,864.3), RMSE (4,578,144), MAE (3,583,768), MPE (3.13), MAPE (7.43), MASE (1.29), and Theil's U (0.44), indicating its superior forecasting accuracy compared to the other two models.

The findings suggest that using the BATS model can enhance the accuracy of forecasting tax revenues, which is crucial for government budget planning and economic stability, as it provides more accurate forecasts for monthly tax revenues in Turkey compared to traditional SARIMA and Random Walk models.

Limitations and Gaps in Existing Research

These previous studies acknowledge several limitations and areas for improvement. Some of the key limitations are:

- Model Assumptions and Data Limitations: The study by Hamza Erdoğdu and Recep Yorulmaz (2020) acknowledged that the inherent assumptions of the forecasting models (e.g., stationarity for ARIMA, specific seasonal patterns for BATS) might not fully capture the real-world dynamics of tax revenue data. Additionally, the limited dataset used in their study (January 2006 to December 2018) may not have captured longer-term trends or structural changes in the Turkish economy.
- Lack of Comprehensive Integration of Digital Insights: While previous studies have explored the application of predictive modelling techniques for tax compliance and revenue forecasting, the integration of digital insights and their impact on taxpayer

behavior has not been extensively addressed. The current study aims to fill this gap by systematically incorporating digital knowledge into the predictive modelling framework to enhance the understanding and prediction of taxpayer compliance.

- Sector-specific Analysis: The study by Hamza Erdoğdu and Recep Yorulmaz (2020) focused on overall tax revenue forecasting, without delving into sector-specific analyses (e.g., corporate taxes, VAT, income taxes). Conducting sector-specific tax revenue forecasts could yield more granular insights and improve overall forecasting accuracy, which the current study aims to address.
- Expanding Model Comparisons: While the previous studies have compared the performance of various forecasting models, the current research aims to extend the model comparison by exploring additional advanced techniques, such as machine learning algorithms (e.g., neural networks, random forests), to provide a more comprehensive evaluation of the predictive capabilities.

By addressing these limitations, the reliability and applicability of tax revenue forecasting models can be enhanced, ultimately aiding in more effective fiscal policy and economic planning. To address these limitations, the researchers suggest the following avenues for future research:

- Extending the dataset to include more recent data or longer historical data to better understand long-term trends and improve model robustness.
- Incorporating exogenous variables (e.g., GDP growth, unemployment rates, inflation) to improve the accuracy of tax revenue forecasts.

- Exploring other advanced forecasting models, such as machine learning techniques (e.g., neural networks, random forests), to provide insights into their relative performance compared to traditional time series models.
- Conducting sector-specific tax revenue forecasts (e.g., corporate taxes, VAT, income taxes) to yield more granular insights and improve overall forecasting accuracy.

In this current study emphasis is on incorporating digital insights, such as patterns in electronic filing, online payments and engagement with digital platforms, aligning with the growing recognition of the impact of digitalization on taxpayer behaviour and compliance. The integration of digital knowledge is a key aspect of the research, as it examines how digitalization influences taxpayer compliance, as highlighted by the FIRS's strategic goals and the African Tax Administration Forum's initiatives on tax compliance.

Tax Compliance and Predictive Modelling

The application of predictive modelling techniques in the area of tax compliance has gained significant traction in recent years. As tax authorities strive to enhance revenue collection and ensure fairness within the tax system, the integration of advanced analytics has emerged as a promising approach (Alm, 2019; Devos, 2014). Predictive modelling has emerged as a powerful tool for tax authorities to identify potential non-compliance and prioritize enforcement efforts. By analysing historical tax data, demographic information, and economic indicators, predictive models can estimate the likelihood of individuals or businesses engaging in tax evasion or avoidance (Gupta & Nagadevara, 2007).

The use of predictive modelling techniques in tax compliance has gained traction globally, but their application in the Nigerian context has been limited. A study by Ogembo (2019) explored the use of logistic regression models to predict tax compliance among small and medium enterprises (SMEs) in Nigeria, using data from tax returns and audits. However, the study highlighted the need for more advanced modelling techniques and the integration of additional data sources to improve predictive accuracy.Another study by Onyekwelu and Nwaiwu (2018) also found that factors like income level, education, and perceived fairness of the tax system were significant determinants of tax compliance.

Previous studies have explored various machine learning algorithms, such as decision trees, neural networks, and logistic regression to develop predictive models for tax compliance (Akinobu et al., 2010; Wu et al., 2012). These models were trained on a range of data sources, including tax returns, audit results, and third-party information reports (Alm & Yunus, 2009).One of the key advantages of predictive modelling in tax compliance is the ability to identify the critical drivers of taxpayer behaviour. With the help of sophisticated algorithms, such as XGBoost models, tax authorities can analyze a vast array of data, including taxpayer characteristics, transaction patterns, and socioeconomic factors, to uncover the most influential predictors of compliance (Kirchler et al.,2008) By incorporating these insights into their compliance strategies, tax authorities can develop more targeted interventions and programs that address the specific needs and challenges faced by different taxpayer segments.

While predictive modelling has shown promise in improving tax compliance, many studies have relied on limited data sources, potentially overlooking important factors that influence taxpayer behaviour (Hashimzade et al., 2014). The integration of emerging data sources, such as digital

interaction patterns and socioeconomic indicators has been limited, hindering the development of more comprehensive and accurate predictive models (Engström et al., 2020).The combination of predictive modelling and digital insights also presents opportunities for tax authorities to leverage the wealth of data available to them. By merging traditional tax data with alternative data sources, such as social media activity, transaction patterns, and demographic information, tax authorities can gain a deeper understanding of taxpayer behaviour and tailor their compliance strategies accordingly (Devos, 2014).

However, the successful implementation of predictive modelling in tax compliance requires a robust legal and procedural framework. Tax authorities must ensure that the use of such technologies aligns with existing regulations, taxpayer rights, and data privacy considerations (Alm, 2019). The Finance Act 2021 introduced provisions to enhance the Federal Inland Revenue Service's (FIRS) ability to deploy technology for tax administration, including assessments and information gathering (PwC, 2022). This presents an opportunity to leverage predictive modelling techniques and integrate various data sources to improve tax compliance strategies in Nigeria.

Continuous monitoring and evaluation of the effectiveness of these models are also crucial to ensure their long-term impact and adaptability to changing compliance landscapes.

This research aims to address these limitations by developing advanced predictive models that leverage a wide range of data sources and employing state-of-the-art machine learning algorithms to capture complex patterns and interactions.

The Impact of Digitalization on Tax Compliance

The rapid digitalization of various aspects of life, including financial transactions, communication, and access to information has significant implications on tax compliance.As tax authorities strive to enhance revenue collection and ensure fairness within the tax system, the integration of digital tools and data-driven insights has emerged as a critical component of their strategies (Alm, 2019; Devos, 2014).

Several studies have explored the impact of digital platforms and technologies on tax compliance. For instance, research has investigated the effectiveness of electronic filing systems in improving compliance rates and reducing administrative costs (Hung et al., 2019). Other studies have also examined the role online payment platforms and digital communication channels in fostering taxpayer engagement and compliance (Alm et al., 2016). The integration of digital insights into predictive modelling approaches remains an area for further research. The understanding how taxpayers engage with digital platforms and technologies can inform the design of user-friendly and effective digital solutions for improving tax compliance.

Digitalization has the potential to facilitate tax compliance in Nigeria by improving transparency, reducing transaction costs and providing convenient filing and payment options (Gchâlàb & Lu, 2019). The implementation of e-filing systems, online payment portals, and automated data processing has not only improved the efficiency of tax collection but also enhanced the transparency and accessibility of the tax system for taxpayers (Alm, 2019).

The integration of digital technologies has also facilitated the streamlining of tax administration and compliance processes.These digital advancements has enabled tax authorities to adopt a more proactive and data-driven approach to compliance monitoring and enforcement. By

leveraging real-time data analytics and automated risk assessment tools, authorities can identify potential non-compliance cases more quickly and allocate their resources more effectively (Devos, 2014)

However, the integration of digitalization in tax compliance also presents new challenges and considerations. Tax authorities must ensure that the use of digital technologies and data-driven insights aligns with existing legal and procedural frameworks, as well as with taxpayer rights and data privacy concerns (Kirchler, 2007).The country faces challenges in terms of limited digital infrastructure, low internet penetration, a significant informal economy (Ogembo, 2019) and the lack of sufficient and comprehensive data sources. .

Digitalization can create new opportunities for tax evasion and avoidance, particularly in the context of cross-border transactions and the rise of the digital economy (Olbert & Spengel, 2019). The increasing complexity of digital business models and the challenges in attributing income to specific jurisdictions have posed difficulties for tax authorities in ensuring compliance (Bräutigam et al., 2017).

In conclusion, the impact of digitalization on tax compliance has been profound, enabling tax authorities to leverage data-driven insights, streamline administrative processes, and adopt more proactive and targeted compliance strategies. As the digital landscape continues to evolve, the effective integration of these technologies will be crucial for tax authorities to enhance revenue collection, promote fairness, and address the challenges of non-compliance.

Socioeconomic Factors and Tax Compliance

The examination of socioeconomic and demographic factors and their impact on company income tax compliance is crucial in the Nigerian context. As the tax authorities strive to enhance revenue collection and ensure fairness within the tax system, socioeconomic factors, such as income level, employment status and education have been identified as significant determinants of tax compliance in Nigeria (Ogembo, 2019; Saad, 2014). The country's large informal sector, income inequality, and low tax morale among certain segments of the population pose challenges for tax authorities (Batrancea et al., 2019).

Income Level:

Studies have shown that the income level of companies can significantly influence their tax compliance behavior. Larger and more profitable companies in Nigeria tend to have a greater capacity to comply with their tax obligations but they may also have more opportunities to engage in tax avoidance or evasion strategies (Onyekwelu & Nwaiwu, 2018). Conversely, smaller companies with lower incomes may face challenges in meeting their tax liabilities, leading to non-compliance. Research has also shown that individuals with lower income levels and those engaged in informal employment activities may face greater barriers to tax compliance such as limited access to information and resources (Ogembo, 2019).

Education and Expertise:

The level of education and expertise within a company's management team can also impact tax compliance. Companies with more educated and experienced personnel, who have a better understanding of tax laws and regulations, are more likely to comply voluntarily (Devos, 2014). This is particularly relevant in the Nigerian context where the complexity of the tax system and frequent changes in tax policies can pose challenges for companies.

Demographic Characteristics:

Demographic characteristics such as the age and ownership structure of a company, can also influence tax compliance. Older, more established companies in Nigeria may have a stronger sense of civic duty and a better understanding of the importance of tax compliance, leading to higher compliance rates (Kirchler, 2007). Additionally, the ownership structure of a company, whether it is publicly or privately owned can impact its approach to tax compliance.

Perceived Fairness of the Tax System:
The perceived fairness of the Nigerian tax system is also a crucial determinant that can shape a company's compliance behaviour. Trust in government and perceptions of fairness, shapes taxpayers' attitudes towards payment of taxes and their compliance behaviour, companies that believe the tax administration is fair and equitable are more prone to prompt voluntary compliance, while those who see the system as unfair will be most likely to engage in non-compliance (Onyekwelu & Nwaiwu, 2018).

Legal and Procedural Frameworks, Taxpayer Rights and Data Privacy Concerns in Predictive Modelling

The successful integration of predictive modelling techniques in enhancing tax compliance must be accompanied by a robust legal and procedural framework that addresses taxpayers rights and data privacy concerns.

Tax Laws and Digitalization: Frameworks for Effective Predictive Models

Finance Act 2021 and Tax Digitalization*:*
The Finance Act 2021 introduced provisions that empower the Federal Inland Revenue Service (FIRS) to leverage technology for tax administration, including assessments and information gathering (PwC, 2022). This digitalization effort, particularly through the deployment of the TaxProMax platform, presents opportunities for more collaborative and data-driven approaches to taxpayer engagement and compliance strategies in Nigeria. The TaxProMax platform facilitates electronic filing, payment, and real-time data collection, improving efficiency and

reducing errors. This digital shift allows FIRS to better track taxpayer activities, automate assessments, and implement timely interventions to improve compliance rates.

Key Provisions of the Finance Act

- **Income Thresholds:** The Finance Act sets an income threshold for CIT and VAT at NGN 25 million per annum, exempting businesses below this threshold from filing and payment obligations (Ogundele, 2020).
- **Digital Services Tax:** Introduces taxation on digital services, aligning with global best practices.

Legal and Procedural Frameworks:

Tax authorities must ensure that the use of predictive modeling and other data-driven compliance strategies aligns with the existing legal and regulatory environment. This includes adherence to tax laws, administrative procedures, and taxpayer rights (Alm, 2019).

Taxpayer Rights:

Taxpayers have fundamental rights to be treated fairly and with respect. Predictive modeling must uphold these rights, including the right to privacy, the right to appeal decisions, and the right to be informed about the use of their data (Kirchler, 2007). Transparency and clear communication about data usage are essential.

Data Privacy Concerns:

The increased use of digital data sources and predictive modeling raises significant data privacy concerns. Tax authorities must ensure compliance with data protection regulations and best practices, implementing robust data governance frameworks and security measures to protect sensitive information (Devos, 2014).

Complexity of Nigerian Tax System:

The Nigerian tax system is governed by multiple laws, such as the Companies Income Tax Act (CITA) and the Value Added Tax Act (VATA). Frequent amendments to these laws create confusion and uncertainty for taxpayers, complicating their ability to comply effectively (Ogundele, 2020).

Key Tax Laws Governing Compliance:

- **Companies Income Tax Act (CITA):** Regulates the taxation of corporate profits in Nigeria. Companies are required to file and pay CIT at 30% of their assessable profits, with an income threshold set at NGN 25 million per annum, below which companies are exempt from CIT.
- **Value Added Tax Act (VATA):** Imposes a 7.5% tax on the supply of goods and services, with an exemption for businesses with annual turnovers below NGN 25 million.

Filing and Payment Deadlines:

- **Companies Income Tax (CIT):** Annual tax returns must be filed within six months of the accounting year-end.
- **Value Added Tax (VAT):** Monthly returns must be filed by the 21st day of the month following the transaction month.

Challenges and Solutions

Nigeria faces several challenges in ensuring high levels of tax compliance:

1. Complexity and Frequent Changes in Tax Laws:

Challenge: The myriad of tax laws and frequent amendments create confusion and uncertainty, hindering effective compliance (Ogundele, 2020).

Solution: Simplifying tax laws and providing clear guidelines can help mitigate this issue.

2. Lack of Taxpayer Education:

Challenge: Many taxpayers are not well-informed about their tax obligations.

Solution: Enhancing taxpayer awareness through education and outreach programs can improve compliance rates (Onyeka & Nwankwo, 2016).

3. Prevalence of Tax Evasion and Avoidance:

Challenge: Tax evasion and avoidance practices are widespread.

Solution: Strengthening tax administration and enforcement, and leveraging technology for better monitoring and detection of non-compliance

Continuous Monitoring and Evaluation:

To address these legal, procedural, and data privacy challenges, tax authorities should establish a comprehensive monitoring and evaluation framework. This framework should assess the ongoing effectiveness of the predictive modeling-based compliance strategies, ensure compliance with relevant laws and regulations, and make necessary adjustments to address emerging issues (Onyekwelu & Nwaiwu, 2018).

Collaboration and Stakeholder Engagement:

Effective implementation of data-driven compliance strategies requires collaboration between tax authorities, legal experts, data privacy specialists, and taxpayer representatives. Engaging

with these stakeholders ensures a balanced approach that enhances revenue collection while protecting taxpayer rights and data privacy.

Conclusion

Tax compliance in Nigeria is influenced by complex laws, frequent amendments, and significant data challenges. Addressing these issues through simplified laws, enhanced taxpayer education, strengthened enforcement, and the use of digital technology can significantly improve compliance. By fostering trust in the government's utilization of tax revenues and providing clear guidelines, Nigeria can enhance revenue mobilization and support sustainable economic development. The holistic integration of advanced analytics, robust legal frameworks, and strong data privacy measures is crucial for the successful and sustainable enhancement of tax compliance in Nigeria.

Predictive Modelling, Taxpayer Engagement and Compliance Strategies

The Nigerian government has traditionally relied on enforcement mechanisms such as audits and penalties, taxpayer education and initiatives to promote tax compliance (Alm et al., 2012). However, these strategies have faced limitations in terms of resource constraints, effectiveness and the potential for unintended consequences, such as creating contentious relationships between taxpayers and the tax authorities. Recognizing these challenges, there has been a growing interest in exploring more collaborative and data-driven approaches to taxpayer engagement and compliance strategies.

The integration of predictive modelling into targeted compliance strategies is crucial in addressing the limitations in effective taxpayer engagement and tax compliance in Nigeria. The

application of predictive modelling and data analytics can provide valuable insights for tax authorities in Nigeria, by analysing socioeconomic and demographic factors that influence company income tax compliance such as income level, industry sector, ownership structure, digital footprints and past compliance history, tax authorities can identify high-risk companies and allocate compliance resources and tailor engagement strategies more effectively (Alm, 2019).

Compliance Strategies:

Engaging with the business community is a fundamental aspect of improving company income tax compliance in Nigeria. Tax authorities should strive to establish open and transparent communication channels with companies, fostering a collaborative environment that encourages voluntary compliance (Onyekwelu & Nwaiwu, 2018). This can be achieved through strategies such as taxpayer education and awareness campaigns, providing effective and timely taxpayer assistance and guidance and actively seeking stakeholder consultation and feedback.

Alongside effective taxpayer engagement, tax authorities in Nigeria should implement targeted compliance strategies that address the specific needs and challenges faced by different segments of the business community. These strategies can include:

- Risk-Based Compliance Monitoring: Utilizing predictive modelling and data analytics to identify high-risk companies and allocating compliance resources accordingly (Alm, 2019).
- Tailored Compliance Interventions: Developing customized compliance interventions, such as educational programs, incentives, or enforcement actions, that cater to the

specific needs and characteristics of different company profiles (Onyekwelu & Nwaiwu, 2018).

- Collaborative Compliance Initiatives: Fostering partnerships and collaborative initiatives with other relevant government agencies, industry associations and their professional bodies and other stakeholders to promote a culture of voluntary compliance and facilitate the exchange of data and other best practices (Devos, 2014).

By leveraging these approaches, tax authorities can enhance revenue collection, promote fairness within the tax system and foster a culture of voluntary compliance among the business community.

None the less the effective implementation of data-driven strategies in Nigeria faces challenges, issues related to data quality, privacy concerns and the ethical use of predictive models must be carefully considered (Alm et al., 2016). Additionally, effective stakeholder engagement and collaboration between tax authorities, policymakers and taxpayers is crucial for the successful implementation of these strategies.

Research Gaps and Opportunities

While previous studies have explored various aspects of tax compliance in Nigeria, several gaps and opportunities for further research have been identified:

In Nigeria there is still limited application and integration of advanced predictive modelling techniques with diverse data sources, such as digital interaction patterns and socioeconomic

indicators. By exploring a wider range of variables, researchers can gain a more comprehensive understanding of the key drivers of tax compliance in Nigeria.

There is the need for a more comprehensive understanding of how digitalization influences taxpayer behaviour across various segments and contexts in Nigeria and how these insights can inform the design of user-friendly digital solutions. Investigating the effectiveness of targeted interventions and compliance programs can also be informed by predictive modelling and digital insights.

Analysing the alignment between Nigerian tax compliance frameworks and international best practices through comparative studies with other jurisdictions can help identify areas for improvement and opportunities for reform.

Applying the theory of planned behaviour can provide insights into the psychological and social factors influencing tax compliance in Nigeria thus Exploring the role of taxpayer attitudes, norms and perceived behavioural patterns in shaping compliance behaviour.

By addressing these research gaps and exploring the identified opportunities, researchers and tax authorities in Nigeria can explore more collaborative and data-driven approaches to taxpayer engagement and compliance strategies, ultimately enhancing revenue collection and the fairness of the tax system whilst leveraging the provisions of the Finance Act 2021 to contribute to the development of more effective and evidence-based tax compliance strategies.

Chapter 3: Methodology

The goal of this research is to develop a machine learning model to predict tax compliance, focusing on late compliance cases. This involves understanding the characteristics of compliant and non-compliant taxpayers and building a predictive model that can identify non-compliance with high accuracy and reliability.

This chapter outlines the research methodology employed in this study aim at developing an integrated approach to enhancing tax compliance in Nigeria through the combination of predictive modelling, digital insights, and demographic factors by attempting to provide a detailed description of the research design, data sources, data collection methods, analytical techniques and ethical considerations involved in the study.

Research Design and Approach

The research design for this project on predictive modelling, taxpayer engagement, and compliance strategies for company income tax compliance in Nigeria adopts a mixed-method approach. This integrated framework combines quantitative and qualitative research methods to comprehensively address the research objectives.

The research design follows a sequential explanatory strategy, where the quantitative phase precedes the qualitative phase. The quantitative component involves the development and validation of predictive models using machine learning techniques. This process leverages various data sources, including tax records, digital interaction data, and demographic indicators, to identify the key socioeconomic and demographic factors influencing company income tax compliance behavior in the Nigerian context.

While the qualitative phase focuses on building and evaluating predictive models, the qualitative phase aims to provide a deeper understanding of the findings and gather feedback on the proposed strategies for enhancing tax compliance. This approach allows for a more comprehensive and nuanced understanding of the factors driving company income tax compliance in Nigeria.

Data Sources and Collection

Data Collection Overview

For this project, data was collected from the Federal Inland Revenue Service (FIRS) of Nigeria. The data included historical tax records comprising tax filings, payments, and compliance information across different tax categories like Pay As You Earn (PAYE), Stamp Duty, Companies Income Tax, Value Added Tax (VAT), Personal Income Tax, Capital Gains Tax, etc. These historical records which spanned multiple years from 2013 to 2023 years of assessment provided a longitudinal view of taxpayer behaviour and trends.

The Tax Types in the Dataset Include:

- **CIT (Company Income Tax)**: Tax on company profits at 30% on reveune of N100 million and above, 25% on revenue of N25 million and less than N100 million, 0% on N25 million and below
- **EDT (Education Tax)**: Tax imposed for funding education at 3% of assessable profit
- **NITDEL (National Information Technology Development Levy)**: Levy for funding IT development at 1% of profit before CIT on turnover of N100 million and above
- **NASENI (National Agency for Science and Engineering Infrastructure)**: Levy for funding science and engineering infrastructure at 0.25% of turnover
- **PTF (Petroleum Trust Fund)**: Funded by a levy on petroleum products at 50% for opertaion sunder NNPC,65.75% for non production sharing contracts under 5 years and 85% after and 30% for upstream gas profits
- **VAT (Value Added Tax)**: Tax on goods and services value addition.
- **WHT (Withholding Tax)**: Tax withheld at source from payments varies between 2.5% for constrtcution,5% for other contracts and 10% for other deductions

- **EMTL (Electronic Money Transfer Levy)**: Levy on electronic money transfers. N50 on transfer of N10,000 or more
- **WVAT (Withholding Value Added Tax)**: 7.5% VAT withheld at source.
- **PAYE (Pay As You Earn)**: Tax on employee income ay 7% for first N300,000 to 24% for N3,200,000 and above
- **CGT (Capital Gains Tax)**: Tax on capital gains from asset sales at 10%
- **SD (Stamp Duties)**: Tax on legal documents at0.75% of authorised share capital

Additionally, data related to digital interactions was gathered, such as statistics on electronic filing submissions and online payment transactions. This data was collected to understand how taxpayers engage with digital channels and how it impacts compliance.

Data Sources and Preparation

To ensure a comprehensive and accurate representation of tax compliance behaviour, digital interaction patterns and demographic indicators in Nigeria, the data used this research on predictive modeling, taxpayer engagement, and compliance strategies for company income tax in Nigeria was collected from two primary sources within the Federal Inland Revenue Service (FIRS) database systems.

These datasets were merged into a single dataset named `df` after a series of cleaning and data transformation steps. The two original datasets and their sources are as follows:

1. TaxPro-Max Dataset (df_1): -

The TaxPro-Max database is a digital tax administration system launched by the FIRS in 2020, enabling seamless registration, filing, payment of taxes, and automatic credit of withholding tax and other credits to taxpayers' accounts.It provides a single-view interface for all taxpayer transactions with FIRS.

The data extracted from the TaxPro-Max database named df_1, covers a period of four tax assessment years from 2020 to 2023.

Data Characteristics The dataset includes information on various tax-related variables, such as CIT, EDT, NITDEL, filing dates, and payment dates. In terms of Structure it had 23 columns and 9410 rows.All columns have the object data type, indicating the need for numerical transformation. It also had Significant number of null values, replaced with 0 however Important columns like Filing Date and Payment Date had no null values.

Column	Non-null Count	Data Type
Taxpayer	9410	Object
CIT	3657	Object
EDT	3657	Objects
Nitdel	3657	Object
Naseni	3657	Object
PFT	3657	Object
PaymentGateway	9410	Object
Office	9410	Object
State	9410	Object
region	9410	Object
Segment	9410	Object
Department	9410	Object
Sector	9410	Object
Filling date	9410	Object
Payment Date	9410	Object
VAT	4527	Object
CIT Group	3657	Object
WHT	1104	Object
EMTL	89	Object
WVAT	16	Object
PAYE	10	Object
CGT	5	Object
SD	5	Object

Table 1: Taxpromax data overview

2. Tax Portal Database (`df_2`): -

The Tax Portal database system maintains the longest historical records of taxpayers in Nigeria.

- The data extracted from the Tax Portal database, named df_2, covered a ten year assessment period from 2013 to 2023.

Data Characteristics

The df_2 dataset has 17 columns and 14,764 rows, with some columns having a significant number of null values, such as CIT, EDT, and VAT which were also replaced by 0. The information on important columns such as payment methods, payment dates had no null values. Values within columns are in the correct format (object, datetime, integer)

Tax Portal Dataset (`df_2`) Overview

Column	Non-null Count	Data Type
Payment Date	14764 non-null	Datetime64[ns]
Taxpayer	14764 non-null	Object
Payment Method	14764 non-null	Objects
Access_period	14764 non-null	Object
Tax_office	14764 non-null	Object
Bank_branch	14764 non-null	Object
Payment_service_provider	14764 non-null	Object
Value_added_tax	14764 non-null	int32
Company_income_tax	14764 non-null	int32
Withholding_tax	14764 non-null	int32
Education_tax	14764 non-null	int32
penalties	14764 non-null	int32
Pre-opertaional_levy	14764 non-null	int32
Capital_gains_tax	14764 non-null	int32
interest	14764 non-null	int32
Personal_income_tax	14764 non-null	int32
pay_as_you_earn	14764 non-null	int32

Table 2: Taxportal data overview

Key Data Characteristics considered in Building a Tax Compliance Prediction Algorithm

Historical Tax Records

Tax records, including historical tax returns, taxpayer names, tax types, and compliance information, were obtained from the Federal Inland Revenue Service (FIRS). These data serve as the primary input for developing predictive models and identifying potential non-compliance patterns (Ogundele, 2020)

Digital Interaction Data

To capture digital insights, the study also utilizes data on taxpayer interaction with digital platforms and sources, such as electronic filing systems and payment gateways. This data provides insights into how taxpayers engage with digital technologies and platforms, as well as their online behavior patterns related to tax compliance (PwC, 2022.)

Demographic Information

Demographic information, including state, region, office, and industry sectors was obtained. These data are used to analyze the impact of demographic factors on tax compliance behaviour, helping to identify trends and patterns across different groups(Adedeji & Oboh, 2012).

Temporal Data

Information related to time, such as payment dates, filing dates, and the specific months and days of payment, was collected. This data is critical for determining tax compliance patterns, enabling the prediction algorithm to account for temporal trends and cycles in taxpayer behaviour (Ogundele, 2020).

Contextual Insights and Considerations

The integration of these various data types provides a comprehensive view of taxpayer behavior, which is essential for building an effective compliance prediction algorithm. By analyzing tax records, digital interaction data, demographic information, and temporal data, the algorithm can identify patterns and trends that signal potential non-compliance. This multi-faceted approach ensures a robust and effective compliance prediction model.

Data Analysis Techniques

The study employed a combination of advanced machine learning algorithms and qualitative data analysis techniques to address the research objectives.

Data Preprocessing

The data collected underwent extensive data preparation, preprocessing, and wrangling, including:

- Data Cleaning: Handling missing values, correcting errors and handling outliers
- Formatting and Structuring: Ensuring consistency in data formats and structures across datasets.
- Integration: Merging the historical tax dataset, digital interaction dataset, and demographic dataset into a unified dataset for comprehensive analysis and modeling

Feature Engineering and Selection

Feature engineering and selection are vital steps in developing robust and accurate predictive models. Feature selection identifies the most relevant and informative features from the data. Feature engineering then creates new variables from the existing dataset, thereby capturing

additional information and uncovering relationships that might not be immediately apparent. Together, these processes enhance the model's ability to understand and predict complex patterns within the data.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical phase in any data science project. In this study, EDA was employed to:

- Summarize Data Characteristics: Using statistical summaries to understand data distribution and central tendencies.
- Identify Patterns and Anomalies: Detecting trends, patterns, and outliers in the data.
- Inform Modeling Decisions: Guiding feature engineering, model selection, and hyperparameter tuning.

Key insights gained from EDA informed the development of a robust and accurate predictive model for tax compliance.

Visualisation Techniques

Data visualization is a powerful tool in exploratory analysis, it allows for the pictorial representation of complex data patterns and relationships. In this study, various visualization techniques were employed to explore the tax compliance dataset. These techniques included:

Count Plots:

Count Plots were used to visualize the frequency distribution of categorical variables, such as the different payment gateways used by taxpayers. This helped in identifying the most popular payment methods and understanding their association with tax compliance.

Histograms with KDE curves:
Histograms with kernel density estimation (KDE) curves were used to visualize the distribution of numerical variables such as company income tax, value added tax and education tax paid. The histogram shows the frequency of data points in each bin, while the KDE curve provides a smoothed estimate of the distribution. These plots provided insights into the shape of the distributions including skewness, modality and potential outliers.

Heatmaps:
Heatmaps were used to visualize the correlation matrix between all numerical variables. This provided a comprehensive overview of the relationships between different variables and helped in identifying potential multicollinearity issues.

Key Terms and Concepts:

Modality:
Modality refers to the number of modes or peaks in the distribution of a variable. The modality of a distribution can provide insights into the underlying patterns and characteristics of the data.

I. Unimodal distribution: A distribution with a single mode or peak.

II. Bimodal distribution: A distribution with two distinct modes or peaks.

III. Multimodal distribution: A distribution with more than two modes or peaks.

Understanding the modality helps in identifying the presence of multiple groups within the data.

Skewness:

Skewness is a measure of the asymmetry of a distribution. It describes the degree and direction of the asymmetry of a distribution around its mean.

Positive skewness: The distribution has a longer right tail, with the bulk of the data concentrated on the left side of the distribution.

Negative skewness: The distribution has a longer left tail, with the bulk of the data concentrated on the right side of the distribution.

Zero skewness: The distribution is symmetric, with the mean, median, and mode aligned.

Skewness can be calculated using the following formula:

$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{(x-\bar{x})}{s}\right)^3$$

Figure 2: Skewness Formula

where:

'*n*' is the number of observations,

'*xi*' the *i*-th observation,

'$\bar{x}$' is the mean, and '*s*' is the standard deviation.

Skewness provides information about the shape and asymmetry of the distribution, which can be useful in understanding the underlying characteristics of the data.

Correlation Matrix:

A correlation matrix is a table that displays the correlation coefficients between all pairs of variables in a dataset. The correlation coefficient measures the strength and direction of the linear relationship between two variables.

The correlation matrix can be visualized using a heatmap, where the values of the correlation coefficients are represented by colors. This provides a comprehensive overview of the relationships between the variables and can help identify potential multicollinearity issues.

The correlation coefficient denoted as r, can be calculated using the following formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Figure 3: Formula Correlation coefficient

Where:- x and y are the two variables

$\bar{x}$ and $\bar{y}$ are the means of the respective variables

Σ represents the sum of the products of the deviations from the means

The correlation coefficient ranges from -1 to 1, where -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship.

Potential Outliers:

Outliers are data points that are significantly different from the rest of the data. They can be detected using various methods like IQR, Z-score, or visual inspections through plots. Potential outliers can be identified whilst using visual techniques such as histograms with KDE curves, where they may appear as observations that are far from the main body of the distribution. They have a significant impact on the analysis and should be carefully examined.

By comprehending these key terms and concepts, analysis and interpretation of the tax compliance dataset can be more effective in identifying significant patterns and in developing well-informed compliance strategies. Employing visualization techniques such as count plots, histograms with KDE curves, and heatmaps enhances the comprehensive understanding of the dataset, allowing for more accurate predictive models and strategic insights.

In conclusion these visualization techniques and statistical concepts are essential for understanding the underlying patterns in the tax compliance dataset, aiding in the development of accurate predictive models.These tools collectively offer a thorough understanding of the tax compliance data, facilitating the identification of trends, correlations, and potential anomalies within the dataset. This, in turn, aids in creating robust models and strategies to improve tax compliance and enforcement.

Model Development and Selection

The study explored and evaluated a range of machine learning algorithms, after evaluating the machine learning algorithms such as Decision Trees, Random Forest and XGBoost the most appropriate algorithm was selected based on factors like predictive performance, computational efficiency, and model interpretability.

These algorithms were trained on the tax records, digital interaction data, and demographic data to identify patterns and make predictions about taxpayer compliance.

Machine Learning Algorithms

To develop advanced predictive models for forecasting taxpayer compliance, gradient boosting, Decision Trees and Random Forest algorithms were employed in this project. These powerful machine learning techniques have demonstrated exceptional performance in various predictive modelling tasks and were well-suited for the complex and multidimensional data utilised in this project.

Decision Trees

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes Simple and interpretable models useful for initial analysis.(IBM)

How It Works:

Decision trees are a widely used supervised learning algorithm that can handle both classification and regression problems. The core principle behind decision trees is the recursive partitioning of the data based on the values of input features, resulting in a tree-like model of decisions.

The decision tree structure begins with a root node, which represents a yes/no condition. This root node then branches out into two child nodes, each corresponding to a different state of the condition. The process continues, with each child node potentially becoming the root of a new subtree or a terminal leaf node.

The depth of a decision tree refers to the maximum number of steps required to travel from the root to the leaves. Deeper trees can capture more complex relationships in the data, but they also risk overfitting. The example decision tree provided below has a depth of 3 and 5 leaf nodes, demonstrating the hierarchical nature of this algorithm.

The key characteristics of decision trees include their intuitive and interpretable structure, the ability to handle both numerical and categorical variables, and their robustness to outliers and missing data. Decision trees are a popular choice for various applications, such as customer segmentation, risk assessment, and predictive maintenance, due to their flexibility and ease of implementation.

Root Node: Start with the entire dataset as the root node.

Splitting: Choose the best feature to split the data based on a criterion like Gini impurity or information gain (for classification) or mean squared error (for regression).

Creating Branches: Split the dataset into subsets based on the chosen feature's values.

Recursive Splitting: Recursively apply the splitting process to each subset until stopping criteria (like maximum depth or minimum samples per leaf) are met.

Leaf Nodes: The final subsets (leaf nodes) contain the predicted values or classes.

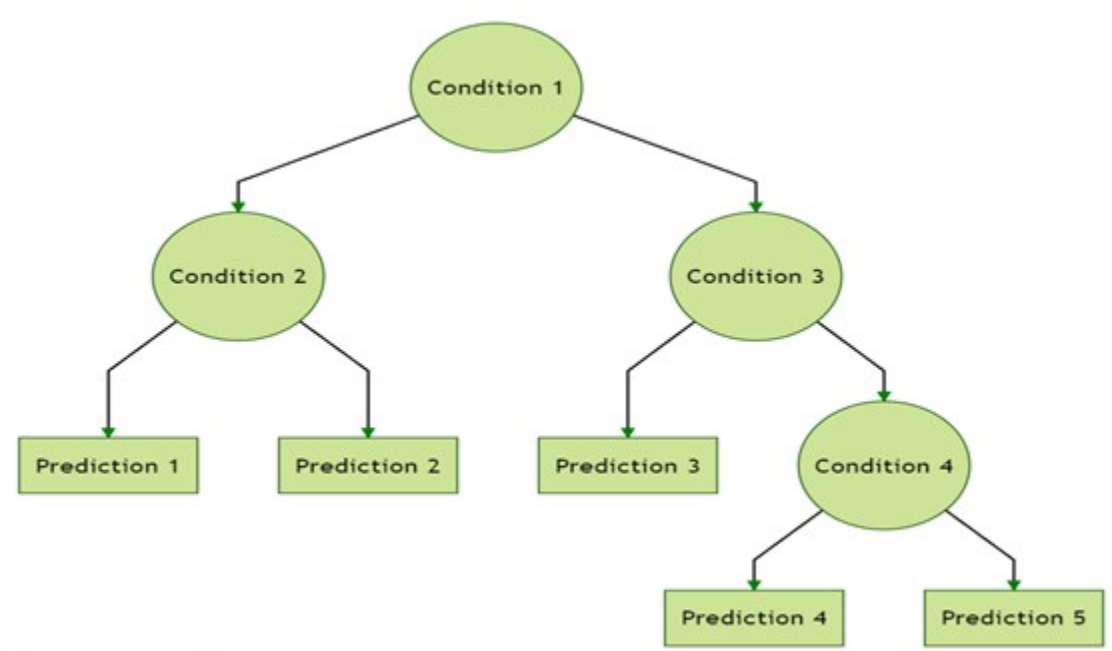


Figure 4: Decision Tree

Decision Trees are simple and interpretable, making them useful for identifying significant factors influencing tax compliance providing clear decision rules that can be easily interpreted by tax authorities. Used in experimental studies to determine key compliance factors, helping in formulating targeted tax policies (Alm & McClelland, 1984).However, they can be prone to overfitting, especially with noisy data.

Random Forests:

Random Forest is an ensemble learning technique that combines multiple decision trees to enhance the overall predictive performance and robustness of the model. The algorithm constructs many decision trees, each trained on a random subset of the features and a random subset of the training data. The final prediction is made by aggregating the predictions of the individual trees, either through majority voting (for classification) or averaging (for regression). The core principle behind the Random Forest algorithm is to leverage the power of multiple decision trees to improve the model's accuracy and stability. By training each tree on a random subset of the features and data, the algorithm introduces diversity and reduces the risk of overfitting, which is a common issue with individual decision trees. The process begins by creating many decision trees, each with its own unique structure and decision rules. During the prediction phase, the algorithm collects the outputs from all the individual trees and combines them to determine the result.

For classification tasks, the algorithm uses majority voting to select the most common prediction among the trees, while for regression tasks, it averages the predictions to obtain the final output. The Random Forest approach harnesses the strengths of decision trees, such as their ability to handle both numerical and categorical variables, their robustness to outliers, and their interpretability. By aggregating the predictions of multiple trees, the algorithm mitigates the weaknesses of individual decision trees, such as their tendency to overfit the training data, and enhances the overall model's performance and generalization capabilities. The versatility and effectiveness of Random Forest have made it a popular choice for a wide range of

applications, including customer segmentation, risk assessment, and predictive maintenance, where the need for accurate and robust predictions is paramount.

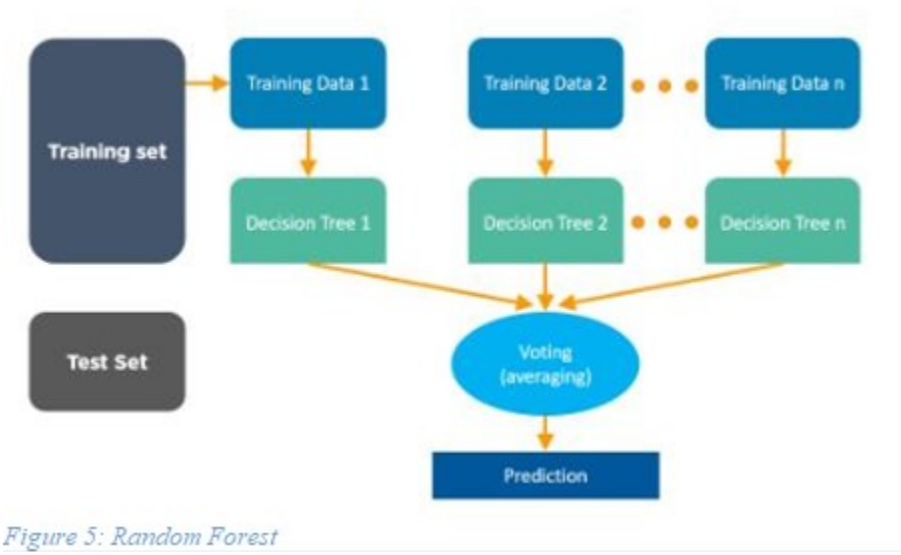


Figure 5: Random Forest

How it Works

Bootstrap Sampling: Create multiple subsets of the original dataset by sampling with replacement.

Building Trees: For each subset, build a decision tree using a random subset of features at each split.

Aggregation: Aggregate the predictions of all the decision trees (e.g., by majority voting for classification or averaging for regression).

Random Forest reduces the variance of individual decision trees by averaging multiple trees, thereby improving accuracy and prunes overfitting. It can handle a large number of features and is effective for both classification and regression tasks.

Random Forest is effective in dealing with large, high-dimensional datasets typical in tax records, providing robust predictions by averaging multiple decision trees

Random Forest: Proven effective in handling large datasets of small business taxpayers, improving the accuracy of compliance predictions and helping tax authorities focus on high-risk cases (Devos & Zackrisson, 2015).

Gradient Boosting Trees:

Are Highly effective for structured data, capable of capturing complex patterns.

XGBoost:

XGBoost, short for Extreme Gradient Boosting is an optimized gradient boosting algorithm known for high performance, scalability and efficiency. It has gained widespread popularity in the machine learning community due to its exceptional performance across a wide range of applications, including tax compliance prediction. XGBoost was utilized in 17 of the 29 winnings Kaggle projects because it is faster and performs better than other machine learning algorithms. (Belle et all 2023)

How XGBoost Works:

XGBoost operates as an ensemble learning technique that integrates numerous weak decision tree models to establish a robust predictive model. The algorithm functions through iterative construction of a series of decision trees, with each subsequent tree aiming to rectify errors made by its predecessors. This iterative process, known as "boosting," enhances the model's predictive performance until it meets a predefined stopping criterion or reaches a maximum number of trees.

Mathematical Formulation

The basic algorithm for a single tree in XGBoost is as follows:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i)$$

Figure 6: XGBoost Formula

Where

ŷ_i: represents the predicted output for the i-th data point in dataset.

Σ from t=1 to T: This is a summation sign indicating that the results are added up from t=1 to T, where T is the total number of functions or models.

f_t(x_i): This represents the output of a function (or model) applied to the i-th data point at iteration t.

The formula is central to how XGBoost, a gradient boosting algorithm, makes predictions.

1. **Initialization:** XGBoost starts with an initial prediction (usually 0.5 for binary classification) for all instances.
2. **Iterative Boosting**: In each iteration, XGBoost adds a new model (tree) that best reduces the loss (error) from the previous step.
3. **Function Output (f_t(x_i))**: Each tree (weak learner) produces a score or output for each instance based on the features (x_i).
4. **Summation (Σ)**: The scores from all trees are summed up to update the predictions iteratively. The equation you provided, ŷ_i = Σ from t=1 to T of f_t(x_i), represents this process where ŷ_i is the updated prediction after adding T trees.

5. **Objective Minimization**: XGBoost optimizes an objective function that includes both the loss and regularization terms to prevent overfitting.
6. **Final Prediction**: After adding all T trees, the final prediction is obtained, which is more accurate than any individual tree's prediction.

In summary, XGBoost builds upon the idea of ensemble learning by combining the outputs of many weak learners (trees) to make a strong final prediction.

The key features that contribute to the success of XGBoost include:

1. **Regularization**: XGBoost integrates Lasso and Ridge Regression techniques to mitigate overfitting and enhance the model's ability to generalize to unseen data..
2. **Parallelization and Cache block**: Although XGBoost cannot train multiple trees concurrently, it optimizes performance by parallelizing the generation of different tree nodes. Data is stored in compressed column format and sorted by feature values within blocks, minimizing sorting costs and computational overheads associated with parallel computation.
3. **Tree Pruning**: Utilizing the max_depth parameter, XGBoost determines the stopping criteria for branch splits and prunes trees in a backward manner. This depth-first approach significantly enhances computational efficiency and yields optimized tree structures.
4. **Cache-Awareness and Out-of-score computation**: XGBoost maximizes hardware resources by implementing cache-awareness strategies, allocating internal buffers for gradient statistics within each thread. It supports out-of-core computation to handle

large datasets that exceed memory capacity, employing data compression techniques to minimize disk space usage.

5. **Sparsity Awareness**: Built-in mechanisms in XGBoost adeptly manage missing values in datasets, accommodating various sparsity patterns without requiring extensive preprocessing steps. Its specialized split finding algorithm further enhances handling of sparse data.
6. **Weighted Quantile Sketch**: The algorithm incorporates a distributed weighted quantile sketch algorithm to effectively identify optimal split points in weighted datasets, enhancing decision-making precision.
7. **Cross-validation**: XGBoost includes built-in cross-validation capabilities, facilitating model validation without external libraries. This feature enables early stopping to prevent overfitting, particularly beneficial for smaller datasets.
8. **Feature Importance**: XGBoost provides mechanisms to evaluate the importance of each feature within the model. This feature is instrumental in feature selection and understanding the factors influencing tax compliance behavior.
9. **Scalability and Speed**: XGBoost is renowned for its scalability and efficient execution speed, outperforming many other algorithms. Its implementation of gradient boosting enhances machine learning development, making it suitable for diverse applications.
10. **Hyper-parameter Tuning**: The algorithm supports automatic adjustment of numerous parameters, enhancing model performance through optimized parameter settings tailored to specific datasets and tasks.

Model Evaluation and Validation

The predictive models were rigorously evaluated and validated using appropriate techniques, such as:

Cross-Validation: Splitting data into training and validation sets multiple times to ensure model robustness.

Holdout Sets: Using a separate dataset for final validation.

Performance Metrics: Evaluating models using precision, recall, F1-score, and the area under the receiver operating characteristic (ROC) curve.

Metric	Description
Precision	The ratio of true positive predictions to all positive predictions
Recall	The ratio of true positive predictions to all actual positives.
F1-Score	The harmonic mean of precision and recall
ROC AUC	Measures the area under the ROC curve, indicating overall model performance

Table 3: Evaluation Metrics and Definitions

Precision

Definition:Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It answers the question: "Of all the positive predictions made, how many were actually correct?"

Formula :

$$Precison=TPTP+FP$$

Recall

Definition:

Recall is the ratio of correctly predicted positive observations to all the observations in the actual class. It answers the question: "Of all the actual positives, how many were correctly predicted?"

Formula:

$$Recall=TPTP+FN$$

F1-Score

Definition:

The F1-Score is the harmonic mean of Precision and Recall. It combines Precision and Recall into a single metric by considering both the false positives and false negatives. It is particularly useful when you need a balance between Precision and Recall.

Formula:

$$F1-Score=2*PrecisionxRecallPrecision+Recall$$

ROC AUC (Receiver Operating Characteristic Area Under the Curve)

Definition:

The ROC AUC measures the area under the ROC curve, which plots the True Positive Rate (Recall) against the False Positive Rate (FPR) at various threshold settings. The AUC value ranges from 0 to 1, with 1 indicating perfect performance.

True Positive Rate (Recall): $TPR = \frac{TP}{TP+FN}$

False Positive Rate: $FPR = \frac{FP}{FP+TN}$

The ROC AUC score is calculated by integrating the ROC curve. The higher the AUC, the better the model is at distinguishing between positive and negative classes.

Accuracy

Definition:

Accuracy is the ratio of correctly predicted observations to the total observations. It measures the overall effectiveness of a model in correctly predicting both positive and negative classes. Accuracy is useful when the classes are balanced, but it can be misleading in the case of imbalanced datasets.

Formula: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

Accuracy is a straightforward metric that provides a quick overview of a model's performance, but it should be used alongside other metrics like Precision, Recall, F1-Score, and ROC AUC, especially when dealing with imbalanced datasets.

Where:

TP = True Positives (correctly predicted positive observations)

TN = True Negatives (correctly predicted negative observations)

FP = False Positives (incorrectly predicted positive observations)

FN = False Negatives (incorrectly predicted negative observations)

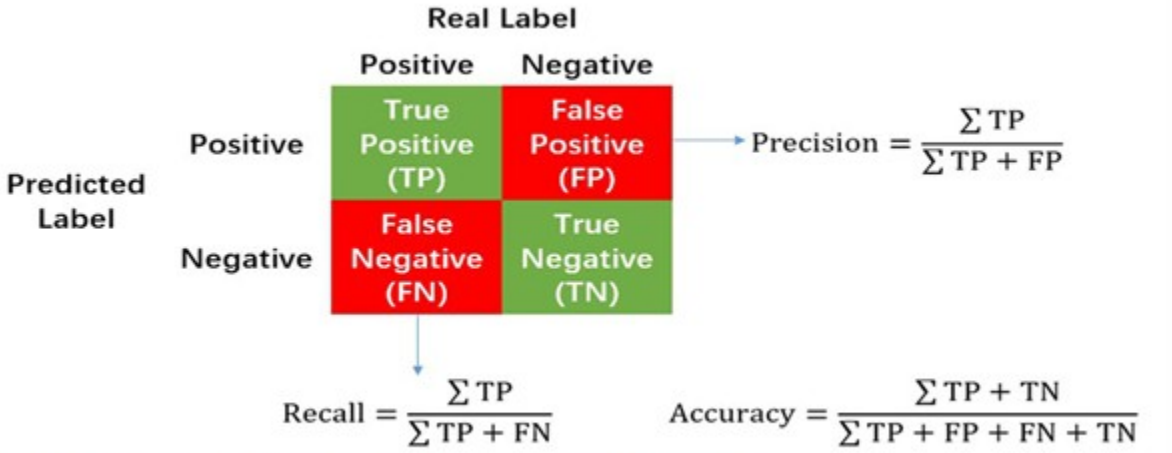


Figure 13: Calculation of Precision, Recall and Accuracy in the confusion matrix.

These metrics are essential for evaluating the performance of classification models, providing insights into the accuracy, completeness, and overall effectiveness of the predictions, ensuring the reliability and generalization of the models.

Hyperparameter Tuning for Gradient Boosting Algorithm

Gradient boosting is a powerful machine learning technique that can be highly effective in predicting tax compliance behavior. However, the performance of the gradient boosting model is heavily dependent on the selection of appropriate hyperparameters. Hyperparameters are the parameters that are set before the training process begins and can significantly impact the model's performance. To optimise the performance of the selected gradient boosting algorithm, a comprehensive hyperparameter tuning process was undertaken. This process involved systematically evaluating different combinations of hyperparameters to identify the configuration that yielded the best predictive accuracy on the validation dataset

Some of the key hyperparameters that were tuned include:

Learning Rate (η):
The learning rate determines the step size at which the model learns from the training data.

- A smaller learning rate can lead to slower convergence but may result in a more stable and generalized model.
- A larger learning rate can lead to faster convergence but may result in a less stable model.
- The optimal learning rate was determined through a grid search or random search approach, where multiple values were tested, and the best-performing rate was selected.

Tree Depth (max_depth):
The tree depth determines the maximum depth of the decision trees used in the gradient boosting model.

- Deeper trees can capture more complex relationships in the data but may be prone to overfitting.
- Shallower trees may be less prone to overfitting but may not capture the nuances in the data.
  The optimal tree depth was identified through a similar grid or random search process.

Regularization Parameters:
Regularization techniques, such as L1 (Lasso) or L2 (Ridge) regularization, can be used to prevent overfitting and improve the model's generalization.

The regularization parameters, such as the regularization strength (α), were tuned to find the right balance between model complexity and generalization.

Other Hyperparameters:
Additional hyperparameters, such as the number of trees also known as number of estimators (n_estimators), the minimum number of samples required to split a node (min_samples_split), and the minimum number of samples required at a leaf node (min_samples_leaf), were also tuned to optimize the model's performance.

Hyperparameter tuning is essential because it directly impacts the performance of the model.

The learning rate controls the step size in each iteration, a smaller learning rate makes the

model training slower but can converge to a better solution. The number of trees

(n_estimators) affects the number of boosting stages, too many trees can lead to overfitting,

while too few may underfit the model. Tree depth controls the complexity of each tree; deeper

trees can capture more complex patterns but also risk overfitting. Regularization parameters like `subsample`, `min_samples_split`, and `min_samples_leaf` help in reducing overfitting by limiting the tree growth.

By systematically evaluating these parameters, the optimal configuration that maximizes the model's accuracy and generalizes well to unseen data can be identified. This process is crucial for building robust and reliable predictive models, especially in complex datasets like tax compliance data where patterns can be subtle and multifaceted.

Model Testing and Evaluation

After tuning the hyperparameters, the machine learning models underwent training on the training dataset. Throughout the training phase, the algorithms acquired knowledge of patterns and correlations existing between the input features and the target variable. Subsequently, the performance of the trained model was assessed on the test dataset by employing suitable evaluation metrics, including accuracy, precision, recall, and F1-score.

Methodological Justification

The mixed-methods approach adopted in this study provides a comprehensive and robust framework for addressing the research objectives. The quantitative component, involving predictive modelling and data analysis, enables the identification of patterns and trends in tax compliance behaviour, as well as the development of accurate predictive models. The integration of diverse data sources, including tax records, digital interaction data, and demographic records ensures a holistic understanding of the factors influencing tax compliance.

The selected data analysis methods which includes sophisticated machine learning algorithms and qualitative data analysis techniques, are well-matched for achieving the research objectives. Machine learning algorithms facilitate the detection of intricate patterns and connections within the various data sets, while qualitative data analysis methods offer insights into the fundamental motivations, perceptions, and experiences associated with tax compliance.

Furthermore, the ethical considerations employed in this section guarantees that the research is carried out in a conscientious and ethical manner, safeguarding the rights and privacy of research participants, as well as the confidentiality of sensitive data.

In conclusion, the methodological decisions taken in this research are intended to offer a rigorous and comprehensive approach to formulating an integrated strategy for improving tax compliance in Nigeria, harnessing the potential of predictive modelling, digital insights, and demographic factors.

Ethical Considerations

This study adheres to strict ethical principles and guidelines to ensure the protection of research participants' rights, data privacy, and confidentiality. Sensitive records such as the tax identification numbers and taxpayer names were dropped from the original data sourced.

Informed Consent

Informed consent was obtained from the tax officials and policymakers prior to the release of the data on tax records. Participants who participated in interviews and focus group discussions were provided with detailed information about the study and the measures taken to ensure data privacy and confidentiality.

Data Privacy and Confidentiality

Strict data privacy and confidentiality measures were implemented to protect the sensitive information obtained from tax records, digital interaction data sources. The data were anonymized to prevent the identification of individual taxpayers. Access to the data was restricted to authorized researchers involved in the study, ensuring that only personnel with a legitimate need to access the data could do so. All data are securely stored using encryption and other security measures and handled in compliance with relevant data protection regulations, such as the General Data Protection Regulation (GDPR) and other local laws.

Ethical Review and Approval

The study proposal, including the methodological approach and ethical considerations, was submitted to the institution's ethics committee for review and approval. This ensures that the study complies with ethical standards and that the rights and welfare of participants are safeguarded. The ethical review process includes an assessment of the potential risks and benefits of the study, the adequacy of the informed consent process, and the measures in place to protect participant confidentiality and data security.

Chapter 4: Data Analysis and Results

This chapter presents the data analysis process and presents the results obtained from the incorporation of predictive modelling techniques, digital insights and demographic indicators within the context of tax compliance in Nigeria. The data pre-processing steps, the analysis procedures and a detailed interpretation of the findings in relation to the research objectives and questions are provided in this chapter.

Data Preprocessing

Before conducting the data analysis, several pre-processing steps were undertaken. The data pre-processing phase involved a series of steps to ensure the data was clean, consistent, and ready for analysis. These steps were crucial for maintaining data integrity and improving the accuracy of the subsequent analyses.

Data Cleaning

Ensuring the quality and integrity of the dataset is a crucial step in any data analysis project. In the context of the tax compliance study, various data cleaning and transformation steps were meticulously followed to prepare the dataset for machine learning. These steps ensured the data was accurate, consistent, and suitable for analysis.

Handling Missing Values

The dataset was examined for missing values using the `isnull().sum()` method. Rows with missing values were identified and filled with 0 for numerical columns. This was necessary as the missing data was represented by "-" and needed to be replaced with a numerical value with `np.nan` using Numpy's method.

Removing Duplicates

Duplicate records can distort analyses and model performance. Thus, duplicate rows were identified and removed using the `drop_duplicates()` method from the pandas library.

Replacing Special Characters

Special characters, such as commas (,), were replaced with an empty string using the `str.replace()` method to prepare the data for conversion to numerical format.

Data Transformation

Data transformation enhances compatibility with machine learning algorithms and improves model performance. Various transformation techniques were applied to ensure compatibility with the chosen machine learning algorithms and improve the predictive performance of the models.

- Columns with incorrect data types were converted to the appropriate data type using the `astype()` method.
- Date columns were converted to datetime format using the `pd.to_datetime()` function.
- New columns were engineered to capture compliance status, such as `vat_compliance` and `cit_compliance`, based on the payment dates (Tabachnick & Fidell, 2013).
- Categorical variables were encoded using techniques like Target Encoding and Label Encoding.
- Continuous numerical variables were scaled or normalized using methods like z-score normalization and Standard Scaling.

Converting Data Types

Columns initially formatted as objects were converted to appropriate numerical data types using the `astype()` method. Date attributes were converted to datetime format using the `pd.to_datetime()` function.

Feature Engineering:

Two new features, *vat_compliance* and *cit_compliance,* were created to capture the compliance status of Value Added Tax (VAT) and Company Income Tax (CIT) payments respectively.

The *vat_compliance* column was created based on the payment date, where payments made on or before the 21st of the month were labelled as *"Non-Default"* and those made after the 21st were labelled as *"Default."*

The *cit_compliance* column was created based on the payment date, where CIT payments made on or before June 30th were labelled as *"Non-Default"* and those made after June 30th were labelled as *"Default"*

Another feature, *"tax_compliance"* was also created to gauge the compliance level of the taxpayers based on existing features.The *"tax_compliance"* column was created to categorize tax payments as *"Early"* or *"Late."* For company income tax, a payment is considered *"Early"* if made on or before June 30th, and *"Late"* if made after. For value-added tax (VAT), a payment is *"Early"* if made on or before the 21st of the month, and *"Late"* otherwise.

Column Renaming:

To improve readability and consistency, column names were converted to lowercase using the str.lower() method.

The "*tin*" column was renamed "*taxpayer*" so it correctly reflects the contents of the column.

Spaces in column names were replaced with underscores using the str.replace() method, following Python's naming conventions for variables.

Data Integration

The study involved two datasets containing relevant information, df_1 from the Taxpromax database and df_2, from the webportal database. To leverage the data from both sources, the datasets were merged based on common columns using the pd.concat() function. The resulting dataset, df, contained relevant information from both df_1 and df_2, enabling a more comprehensive analysis.

Merged Tax Dataset (`df') Overview

Column	Non-null Count	Data Type
Taxpayer	11305	Object
Company_income_tax	11305	float64
Education_Tax	11305	float64
National_information_technology_levy	9396	float64
Naseni	9396	float64
Petroleum_trust_fund	9396	float64
PaymentGateway	9396	object
Office	9396	object
State	9396	object
region	9396	object
Segment	9396	object
Department	9396	object
Sector	9396	object
Filling date	9396	object
Payment Date	11305	object
Value_added_tax	11305	object
CIT_group	9396	float64
Withholding_tax	11305	float64
Electronic_money_transfer_levy	9396	float64
Withholding_value_added_tax	9396	float64
Pay_as_you_earn	11305	float64
Capital_gains_tax	11305	float64
Stamp_duties	9396	float64
vat_compliance	11305	object
cit_compliance	11305	object

Table 1: Table 4: Merged dataset (df)

Dataset Summary: Understanding the Merged Data

The table provided gives an overview of the `df` dataset, which resulted from merging two datasets, `df_1` from the Taxpromax database and `df_2` from the webportal database. This merging was done using the `pd.concat()` function based on common columns.

Insights from the Merged Dataset

The dataset 'df' contains 26 columns with 11,305 entries. including various tax types, taxpayer information and compliance indicators but some columns have only 9396 non-null entries, indicating missing data for those fields.

- The data types of the columns are a mix of 12 numerical columns (float64) and 14 categorical columns (object).
- The Date columns (e.g., Filing Date, Payment Date) had been converted in early preprocessing to facilitate machine learning

The key details of the dataset:

1. Taxpayer: This column contains the taxpayer information as an object data type, with 11,305 non-null values.

2. Company_income_tax, Education_Tax, Withholding_tax, Pay_as_you_earn, Capital_gains_tax: These are numerical columns representing various tax-related payments, all with 11,305 non-null values and a data type of float64.

3. National_information_technology_levy, Naseni, Petroleum_trust_fund, Electronic_money_transfer_levy, Withholding_value_added_tax, Stamp_duties: These are also

numerical columns related to different tax types, but they have 9,396 non-null values, indicating that some data is missing for these features.

4. PaymentGateway, Office, State, region, Segment, Department, Sector, Filling date, Payment Date, Value_added_tax: These are categorical columns stored as object data types, with 9,396 non-null values.

5. vat_compliance and cit_compliance: These are two new columns created to capture the compliance status of Value Added Tax (VAT) and Company Income Tax (CIT) payments, respectively. Both columns have 11,305 non-null values and are of object data type.

6. CIT_group: This column appears to be a numerical feature related to the company income tax group, with 9,396 non-null values and a data type of float64.

The key insights from this dataset are:

It is a merged dataset from two sources (df_1 and df_2) to leverage the data from both sources.

The dataset contains a mix of numerical and categorical features related to tax payments and taxpayer information.

There are some missing values in certain columns, which will need to be addressed during the data preprocessing and modeling stages.

The creation of the vat_compliance and cit_compliance columns suggests an interest in predicting tax compliance behavior.

This dataset provides a comprehensive view of the tax compliance landscape and can be used to develop predictive models to enhance tax compliance strategies.

Exploratory Data Analysis

Before conducting the data analysis, several pre-processing steps were undertaken. The data pre-processing phase involved a series of steps to ensure the data was clean, consistent, and ready for analysis. These steps were crucial for maintaining data integrity and improving the accuracy of the subsequent analyses.

Structure and Summary Statistics

The structure and summary statistics of the dataset were examined using methods such as head(), info(), and describe().These techniques provided an overview of the dataset, including the number of rows and columns, data types, and basic statistical measures (e.g., mean, median, standard deviation).

This initial exploration helped to identify any potential data quality issues or anomalies that needed to be addressed during the data preprocessing stage.

The summary statistics offers detailed insights into the dataset's key attributes. Focusing on the `Taxpayer`, `Company_income_tax`, `Education_Tax`, `PaymentGateway`, and `Payment Date` columns.

Here's an interpretation of the statistics:

`Taxpayer`

Count: 11,305 unique entries.

Unique: There are 2,092 unique taxpayers in the dataset.

Top: The taxpayer with the highest frequency is "Zuma 828".

Frequency (Freq): "Zuma 828" appears 206 times in the dataset.

	Taxpayer	Company_income_tax	Education_Tax	Value_added_tax	PaymentGateway	Payment Date
count	11305	11305	11305	11305	9396	11305
mean	NaN	5.540128e+4	4.364374e+04	1.196914e+5	NaN	2022-05-16
std	NaN	6.178399e+05	2.378331e+05	7.796208e+5	NaN	NaN
min	NaN	NaN	NaN	NaN	NaN	2013-08-02
25%	NaN	NaN	NaN	NaN	NaN	2021-11-23
50%	NaN	NaN	NaN	5.625000e+02	NaN	2022-08-04
75%	NaN	NaN	NaN	5.250000e+04	NaN	2023-05-15
max	NaN	NaN	NaN	1.898402e+07	NaN	2023-12-31
Unique	2092	1000.0	2117.0	3678.0	3	NaN
Top	Zuma 828	0.0	0.0	0.0	Remita	NaN
Freq	206	7302.0	7531.0	5597.0	9053	NaN

Table 2: Table 5: Statistics Summary Of key variables

`Company_income_tax`

Count: 11,305 entries.

Mean: The average company income tax is 55,401.28, this mean value seems quite high, suggesting that there are some large tax payments which are inflating the average

Standard Deviation (std): 617,839.90 is very high relative to the mean this indicates a high variance in tax amounts.

Unique: There are 1,000 unique tax amounts.

Top: The most frequent tax amount is 0.0.

Median (50%):The median value is 0, meaning more than half of the companies in the dataset did not pay any company income tax.

Maximum (max):The maximum value is 34,496,120. This very high value indicates the presence of a few companies with extremely high tax payments, contributing to the high mean and standard deviation

Frequency (Freq): This tax amount (0.0) appears 7,302 times, showing that a large number of entries have zero tax paid.

`Education_Tax`

Count: 11,305 entries.
Mean: The average education tax is 43,643.74.Similar to the company_income_tax, this mean value suggests some high education tax payments
Standard Deviation (std): 237,833.10, indicating significant variability.

Unique: There are 2,117 unique education tax amounts.

Top: The most frequent education tax amount is 0.0.
Median (50%):The median value is 0, meaning more than half of the companies in the dataset did not pay any education tax.

Maximum (max):The maximum value is 9,289,000, indicating the presence of some companies with very high education tax payments

Frequency (Freq): This tax amount (0.0) appears 7,531 times, indicating that many entries have zero education tax paid.

`PaymentGateway`
Count: 9,396 entries.

Unique: There are 3 unique payment gateways.

Top: The most frequently used payment gateway is "Remita".

Frequency (Freq): "Remita" is used 9,053 times, making it the predominant payment gateway.

`Payment Date`
Count: 11,305 entries.

Mean: The average payment date is May 16, 2022.

Standard Deviation (std): Not provided.

Min: The earliest payment date is August 2, 2013.

25% Percentile: November 23, 2021.

Median (50%): August 4, 2022.

75% Percentile: May 15, 2023.

Max: The latest payment date is December 31, 2023.

Overall Insights

1. Skewness: The distributions of both `company_income_tax` and `education_tax` are highly skewed to the right. Most data points are zero, with a few very high values, suggesting that many companies might be exempt or have no tax liability.
2. High Variability: The high standard deviations and the significant differences between the mean and median indicate substantial disparity in tax payments among companies. While some companies pay substantial taxes, many others pay little to none.
3. Outliers: The presence of very high maximum values indicates the existence of outliers. These outliers represent companies with exceptionally high tax payments, likely large corporations with substantial revenues.
4. Concentration at 0: A significant portion of the dataset has zero values for both `company_income_tax` and `education_tax`. This suggests that many companies either fall under exempt categories or have not paid these taxes.
5. Payment Gateway: "Remita" is overwhelmingly the preferred payment gateway, with a much higher frequency than other gateways. This dominance indicates that most taxpayers prefer using this platform for their payments.
6. Payment Dates: The dataset spans a wide range of payment dates, with most payments made within the last few years. This trend highlights the growing impact of digitalization on tax payment processes.

These methods provided an overview of the dataset, including the number of rows and columns, data types, and basic statistical measures (e.g., mean, median, standard deviation).

Categorical Variable Visualization

The distribution of categorical variables, such as Digital Payment Platforms, Most Popular Taxpayers, were visualized using countplots from the Seaborn library. It showed the Remita Payment Platform to be by far the most popular among the taxpayers and Zuma 828 Coal Limited to be the Most Popular Taxpayer.

Count Plot of Payment Platforms

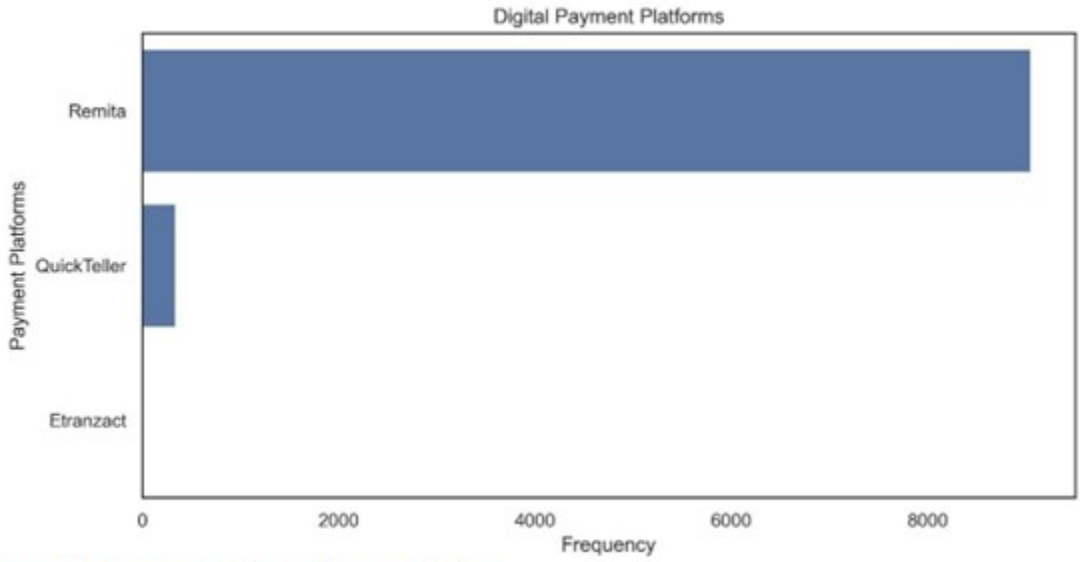


Figure 14: Distribution of Digital Payment Platform

Observation: The count plot for payment gateways indicates that Remita is overwhelmingly the preferred platform, with a frequency count significantly higher than any other gateway.

Implication: This preference might be due to Remita's user-friendly interface, reliability, or wider acceptance among taxpayers.

Count Plot of Industry Sectors

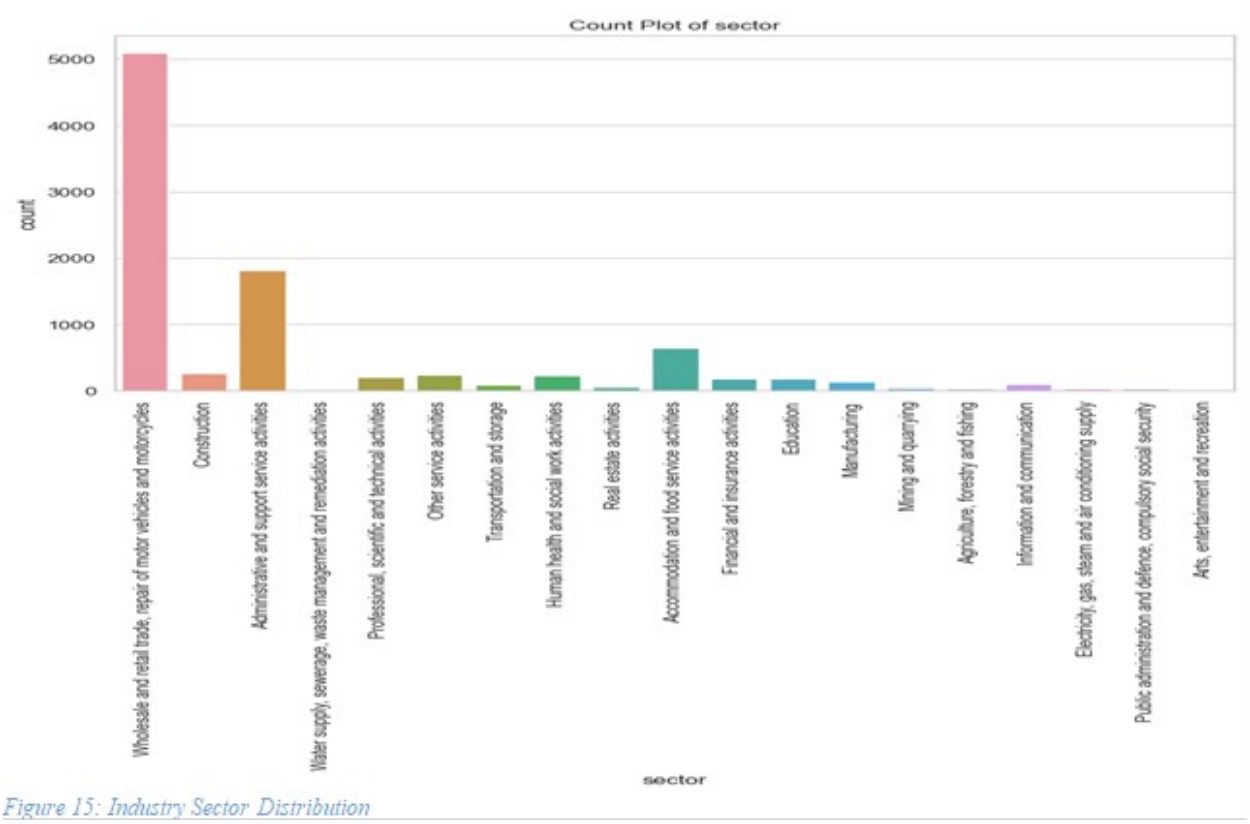


Figure 15: Industry Sector Distribution

Description:

The count plot visualizes the distribution of the various industry sectors in the dataset. Each bar represents the frequency or count of entries belonging to a specific sector.

Insights:

- Dominant Sector: The "Wholesale and retail trade; repair of motor vehicles and motorcycles" sector dominates with the highest count, exceeding 5000 entries. This suggests that a significant portion of the data pertains to businesses involved in retail trade and vehicle repair.
- Other Sectors: Other sectors such as "Construction," "Administrative and support service activities," and "Water supply; sewerage, waste management and remediation activities" also have noticeable counts but are much smaller compared to the dominant sector.
- Diversity: There is a wide range of sectors represented in the dataset, indicating diversity in the types of businesses covered.

Implications:

- Policy Focus: Policymakers might need to pay special attention to the dominant sectors when designing tax policies and compliance strategies.
- Sector-Specific Analysis: Further analysis could be beneficial to understand the tax behaviors and compliance rates within the most represented sectors.

Taxpayer Frequency:

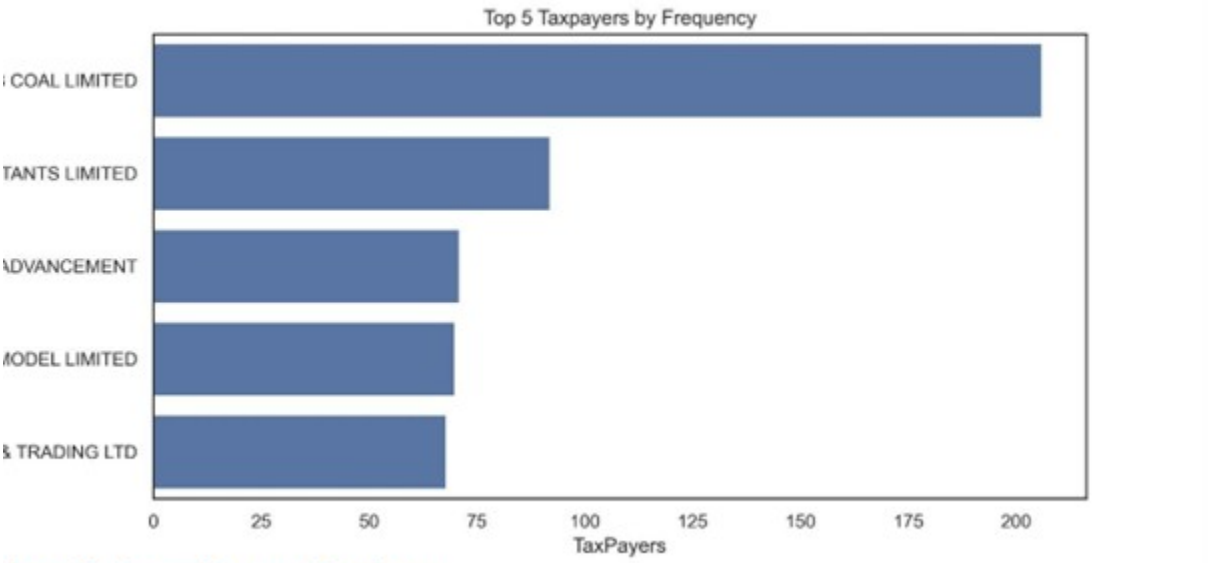


Figure 16: Taxpayer Frequency Distribution

Observation: Among the taxpayers, Zuma 828 Coal Limited stands out with the highest frequency, suggesting that it is a regular and significant contributor to tax payments.

Implication: This entity's prominence in the dataset could be due to its large operations and substantial tax obligations.

Count Plot of CIT Compliance

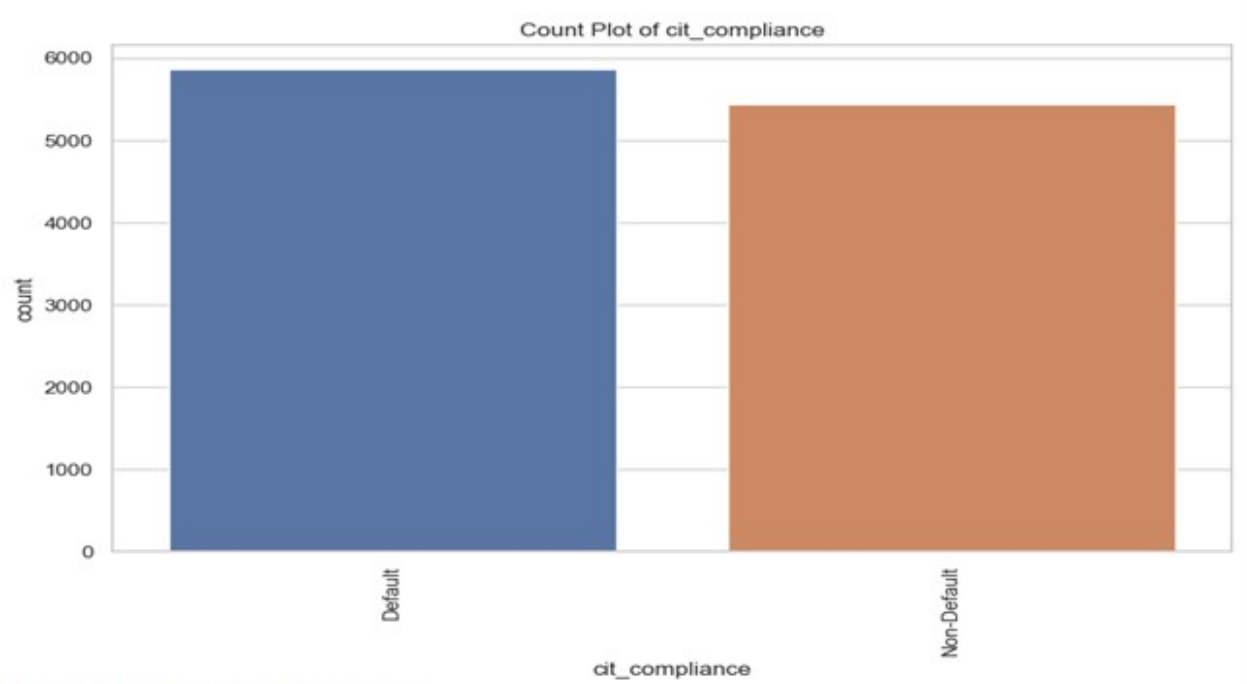


Figure 17: CIT Compliance Distribution

Description

The count plot shows the distribution of Company Income Tax compliance statuses, categorized as "Non-Default" and "Default."

Insights

Majority Non-Default: The majority of entries fall under the "Non-Default" category, indicating that most taxpayers comply with CIT regulations on time.

Significant Default: There is still a considerable number of "Default" entries, highlighting instances of late CIT payments.

Implications

Compliance Strategies: Efforts to reduce the default rate could include stricter enforcement or incentives for early payments.

Further Analysis: Investigating the reasons behind defaults could help in developing targeted interventions to improve compliance.

Count Plot of VAT Compliance

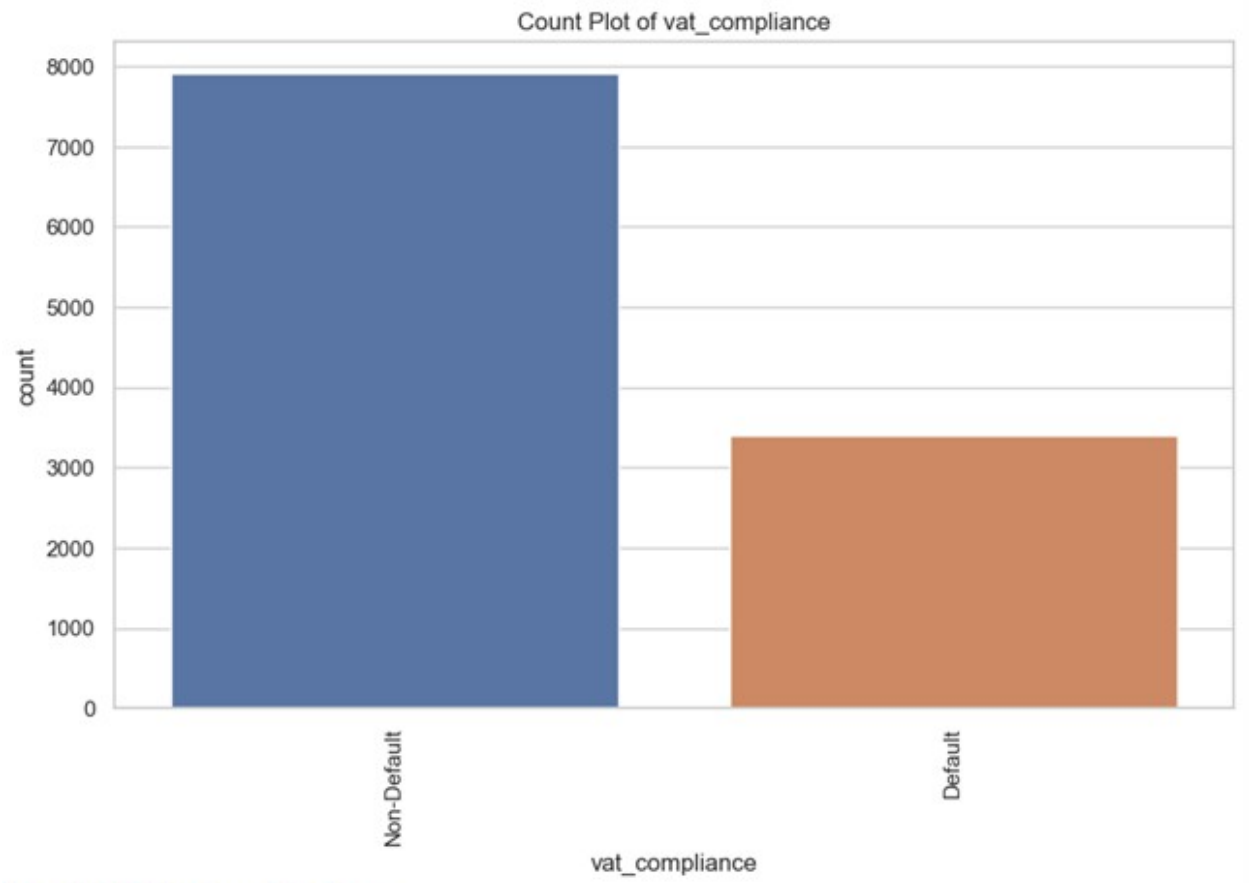


Figure 18: VAT Compliance Distribution

Description:

The count plot shows the distribution of VAT compliance statuses, categorized as "Non-Default" and "Default."

Insights:

Majority Non-Default: The majority of entries fall under the "Non-Default" category, indicating that most taxpayers comply with VAT regulations on time.

Significant Default: There is still a considerable number of "Default" entries, highlighting instances of late VAT payments.

Overall Discussion on CIT and VAT Compliance

Skewness and Variability

Skewed Distributions:
Both company income tax and VAT distributions are highly skewed to the right, with many data points at 0 and a few very high values. This indicates that many companies might be exempt or have no tax liability, while a few pay significantly higher taxes.

High Variability:
The large disparity in tax payments among companies is evident from the high standard deviations and the significant difference between the mean and median values.

Outliers and Concentration

Outliers: The presence of very high maximum values suggests that a few companies have exceptionally high tax payments, which could be due to higher revenues or specific tax liabilities.

Zero Concentration:
A significant portion of the dataset has zero values for both taxes, implying many companies either fall under exempt categories or haven't paid these taxes

Conclusion:
Countplots provide a visual representation of the frequency or count of each unique category, allowing for quick identification of dominant categories and potential outliers.The visualization of categorical variables using count plots provided clear insights into the tax compliance dataset. The dominance of the Remita payment platform, the wholesale and vehicle repair sector and the frequent tax payments by Zuma 828 Coal Limited were highlighted, offering valuable information for understanding taxpayer behavior and preferences. These insights can inform strategies for enhancing tax compliance and optimizing the use of digital payment platforms.

Tax Compliance (Target Variable) Distribution

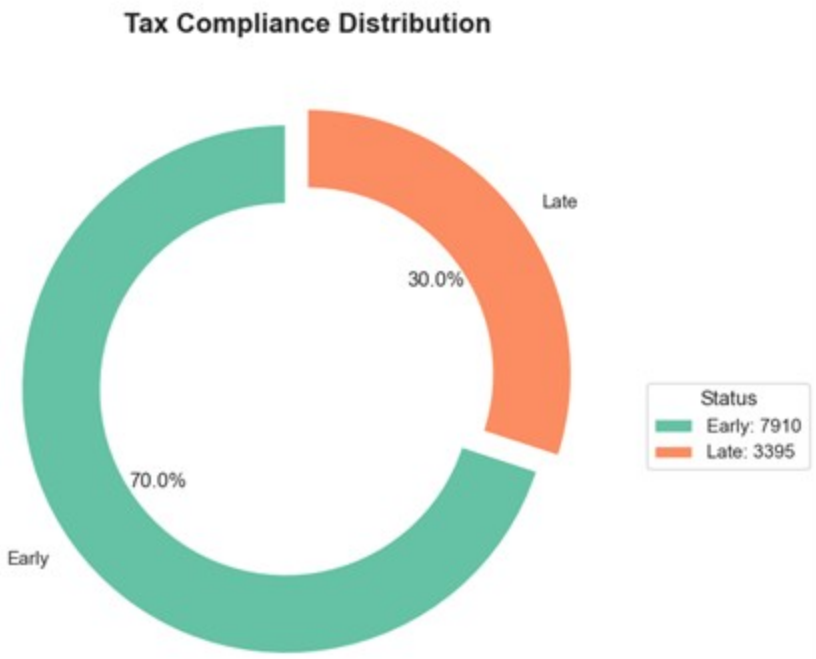


Figure 19: Tax Compliance(Target variable) Distribution

The analysis of the tax compliance distribution, which serves as the target variable for prediction, reveals significant insights into taxpayer behavior. The compliance status is divided into two categories: early-compliant and late-compliant.

Key Findings:

Majority Early-Compliant: The data shows that the majority of taxpayers, approximately 70%, were early-compliant. This indicates that these taxpayers adhered to tax deadlines and regulations, ensuring timely tax payments.

Significant Late-Compliant:The remaining 30% of taxpayers were late-compliant. This group represents taxpayers who failed to meet tax deadlines and paid their taxes late.

Implications:

High Early Compliance Rate: A high rate of early compliance is a positive indicator of effective tax administration and taxpayer cooperation. It suggests that most taxpayers are either motivated or compelled to comply with tax regulations on time.

Focus on Late Compliance: The 30% of late-compliant taxpayers highlight an area that requires attention. Understanding the reasons behind late compliance could help in designing targeted interventions to improve compliance rates.

Numerical Variable Visualization

Distribution Analysis of Numerical Variables

The distributions of numerical variables, such as company income tax, education tax, and withholding tax, were visualized using histograms with kernel density estimation (KDE) curves. These visualizations provide crucial insights into the data's distribution, including skewness, modality, and potential outliers, which can guide subsequent modeling and analysis decisions.

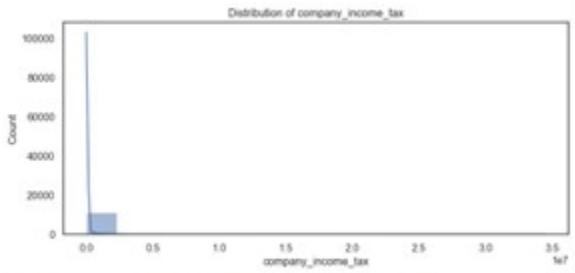


Figure 20: Company Income Tax Distribution

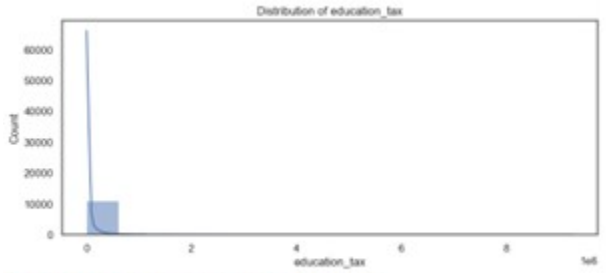


Figure 21: Education Tax Distribution

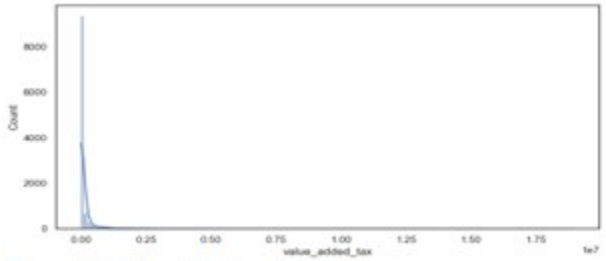


Figure 22: Value added tax distribution

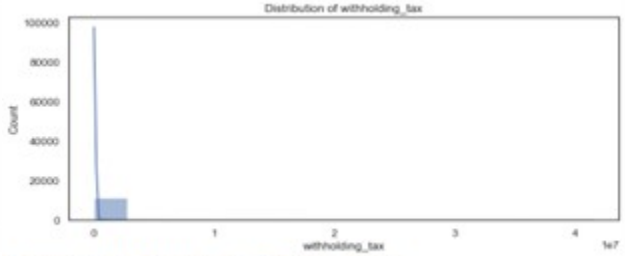


Figure 23: Withholding Tax Distribution

Skewness:

The distributions for company income tax, education tax payments are heavily skewed to the right, indicating that most data points are concentrated towards the lower end of the tax values. This suggests that the majority of companies fall within a lower tax bracket, potentially benefiting from the tax exemption for companies with revenue below ₦25,000,000.

However, the long tails extending towards the right indicate the presence of a few companies with significantly higher tax payments, which could be large corporations with substantial profits. To better visualize the distribution of the company income tax variable, a log transformation was applied using the function `*np.log1p(df['company_income_tax'])*`. This transformation helped to convert the skewed distribution into a more normal distribution, allowing for a clearer representation of the data, including the companies with higher tax liabilities.

Applying Log Transformation

To address the skewness and better visualize the distribution, a log transformation is applied to the company income tax data. This transformation is performed using the np.log1p() function, which applies the natural logarithm to the data after adding one to each value (to handle zero values without issues).

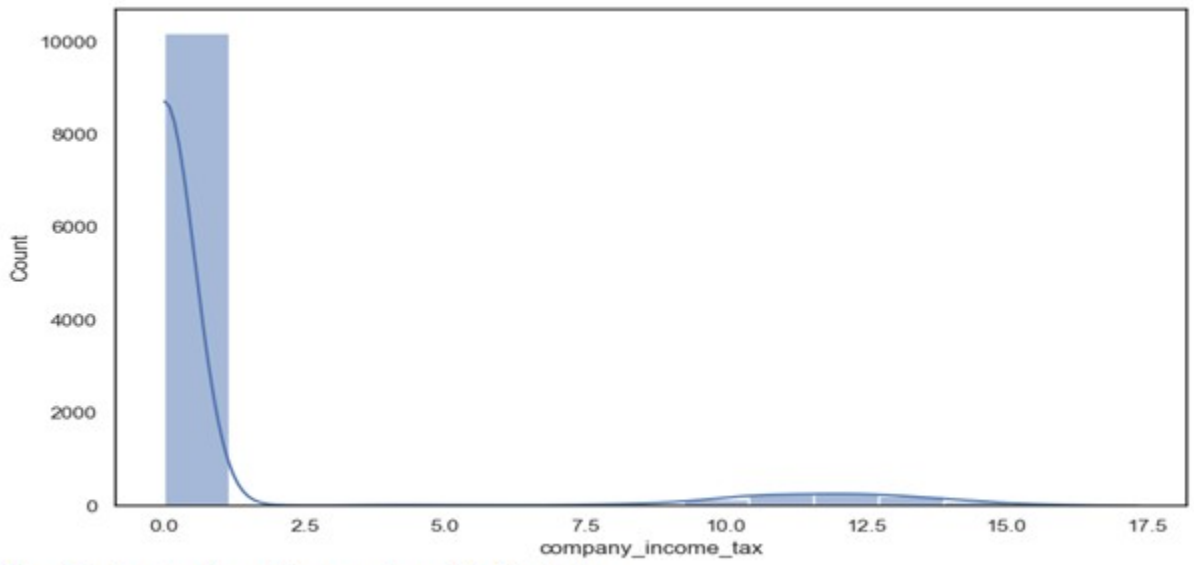


Figure 24: Log transformed Company Income Tax Distribution

Insights from Log Transformed Dataset

- **Normalization**: The log transformation reduces skewness, converting the heavily right-skewed distribution into a more normal distribution. This transformation compresses the range of values, making it easier to visualize and interpret.
- **Visualization**: In the transformed data, the bulk of the tax payments are more evenly distributed. The concentration of lower-end values is less pronounced, and the right tail, which represents companies with high tax payments, is more clearly visible.
- **Insight into High-Paying Companies**: The log-transformed histogram highlights the presence of companies with exceptionally high tax payments. These entities are likely large corporations with substantial profits, which correlate with higher tax liabilities. This clearer visualization aids in identifying and analyzing these outliers more effectively.

Practical Implications

- **Policy Implications**: The transformation and subsequent analysis can help policymakers understand the distribution of tax burdens more clearly. Recognizing the significant contributions from high-paying companies can guide decisions on tax regulations and enforcement strategies.
- **Compliance Monitoring**: Identifying companies with higher tax liabilities through the transformed data can help tax authorities focus their compliance and auditing efforts more effectively.

Conclusion

The log transformation of company income tax data provides a more normalized and interpretable distribution. By reducing skewness, it enhances the visualization and analysis of tax payments, making it easier to identify patterns and outliers. This approach supports better-informed decisions and strategies in tax policy and compliance monitoring

Time Series Analysis

The time series analysis indicates a noticeable increase in tax payments, especially post-2020. This could be attributed to the impact of digitalization and improved tax collection mechanisms. Significant peaks in tax payments highlight periods of high tax collection, possibly linked to specific events or deadlines.

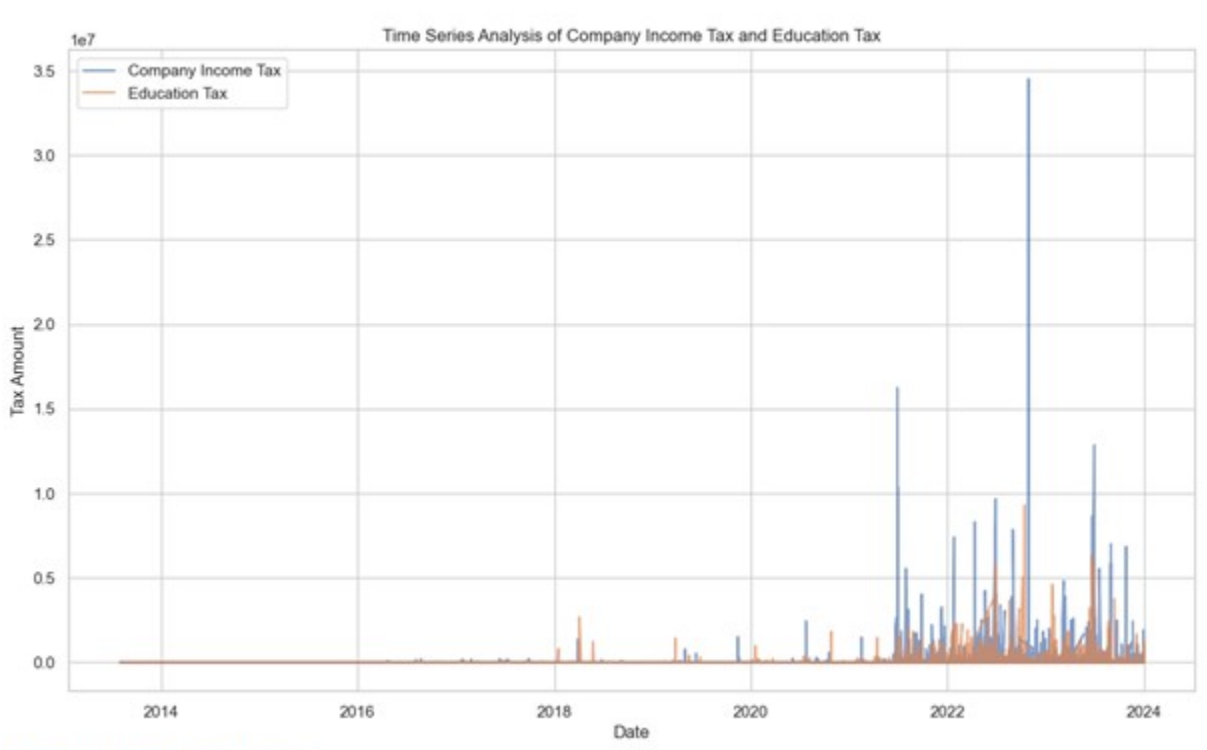


Figure 25: Time Series Analysis

Insights:

Trends Over Time: There is a noticeable increase in both company income tax and education tax payments over recent years, especially post-2020.
Peaks: Significant peaks in tax payments are observed, indicating periods of high tax collection.
Digitalization Impact: The increase in tax payments in recent years could be linked to the impact of digitalization and improved tax collection mechanisms.

Implications:

Seasonality: Further investigation into seasonality or specific events causing peaks could provide insights into tax payment behaviors.
Policy Impact: The trends may reflect the effectiveness of recent policy changes or tax collection initiatives.

Correlation Analysis

A correlation heatmap was used to examine the relationships between numerical features. The majority of correlation coefficients were found to be close to 0, indicating poor correlation among the numerical variables. This suggests that the variables are largely independent of each other, which has implications for the modeling stage, as it indicates that multicollinearity is not a significant concern.

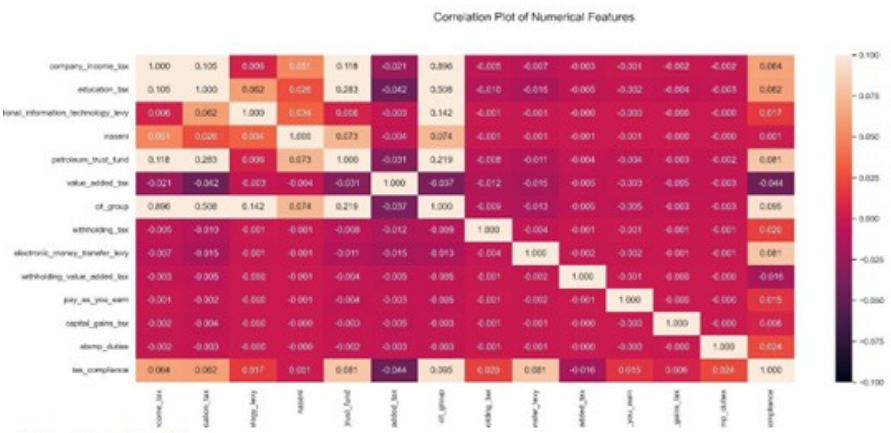


Figure 26: Heatmap of Correlation Coefficient

Findings from the Correlation Heatmap

Low Correlation: The majority of the correlation coefficients were close to 0, indicating a poor correlation among the numerical features. This suggests that the numerical variables, such as company_income_tax, education_tax, and withholding_tax, do not have strong linear relationships with each other

Implications: Poor correlation implies that these tax variables are likely influenced by different factors and may not predict each other well. This could necessitate the use of more complex modeling techniques, such as nonlinear algorithms or ensemble methods, to capture the nuances in the data.

The careful considerations of the implications of the low correlation coefficients observed in the correlation heatmap, informed decisions regarding feature selection, model development, and the overall approach in this tax compliance predictive modeling task.

Machine Learning Model Development

The machine learning model development phase is a critical step in this study, as it aims to create a predictive tool that can accurately forecast taxpayer compliance behaviour. This section details the process of the model's pipeline preprocessing, models selection, training and evaluation. It highlights the rationale behind the chosen approach and the implications of the results.

Creating and Analyzing the Target Variable

The development of the target variable *`tax_compliance`* was crucial for the machine learning model as it represents the compliance behaviour aimed to predict. The *`tax_compliance`* variable was created by first defining a new column *`tax_type`* that categorizes each row as either *'VAT'* or *'CIT'* based on the existence of the *`vat_compliance`* or *`cit_compliance`* columns. This was achieved using the *`np.select()`* function, which allows for conditional assignments based on multiple conditions.

Below is a breakdown of the steps taken to create this variable and the output explained in detail.

Creation of `tax_type` Column
The objective is to categorize each row based on the type of tax compliance (VAT or CIT).
Code Explanation:
Conditions: Two conditions are checked using *`np.select`*:
`df['vat_compliance'].notnull()`: Checks if the *`vat_compliance`* column is not null, indicating the row pertains to VAT.

`df['cit_compliance'].notnull()`: Checks if the `*cit_compliance*` column is not null, indicating the row pertains to CIT.

Values: Corresponding values assigned based on conditions are *`['VAT', 'CIT']`*.

Default: If none of the conditions are met, *`np.nan`* is assigned.

Result:

The `tax_type` column will contain 'VAT' or 'CIT' based on the presence of `vat_compliance` or `cit_compliance` values.

Creation of `tax_compliance` Column

The aim is to categorize each row into 'Early' or 'Late' based on the type of tax and compliance status.

Code Explanation:

Conditions: Four conditions are checked:

`(df['tax_type'] == 'VAT') & (df['vat_compliance'] == 'Non-Default')`: Indicates early VAT compliance.

`(df['tax_type'] == 'VAT') & (df['vat_compliance'] == 'Default')`: Indicates late VAT compliance.

`(df['tax_type'] == 'CIT') & (df['cit_compliance'] == 'Non-Default')`: Indicates early CIT compliance.

`(df['tax_type'] == 'CIT') & (df['cit_compliance'] == 'Default')`: Indicates late CIT compliance.

Values: Assigned values are `['Early', 'Late', 'Early', 'Late']`.

Result:

The *`tax_compliance`* column will contain *'Early' or 'Late'* based on the tax type and compliance status.

Dropping the *`tax_type`* Column

The *`tax_type`* column is no longer needed for analysis and is therefore dropped.

df.drop(columns=['tax_type'], inplace=True)

Printing *`tax_compliance`* Value Counts

Finally, the value counts of the *`tax_compliance`* column are printed to understand the distribution of compliance statuses.

print(df['tax_compliance'].value_counts())

Output:

The output will show the number of rows for each unique value in the `tax_compliance` column:

`Early`: 7910 rows

`Late`: 3395 rows

This distribution indicates that the dataset has a higher number of rows with 'Early' compliance compared to 'Late' compliance. This insight is crucial for understanding the overall compliance behavior and will inform the model development process.

Summary

By creating the *`tax_type`* and *`tax_compliance`* columns and then dropping the unnecessary *`tax_type`* column, we have effectively prepared the target variable for our machine learning model. The value counts provide a clear picture of the compliance behavior distribution, which is essential for model training and evaluation. This structured approach ensures that our dataset is well-prepared for predictive modeling, ultimately aiding in accurate forecasts of taxpayer compliance behavior.

Machine Learning Pipeline Creation

The initial step in building the predictive algorithm is to prepare the dataset for machine learning by creating a preprocessing pipeline using scikit-learn. The use of a pipeline ensures that he necessary data transformations are consistently applied during both the training and prediction phases.

Preprocessing Steps:

Categorical Encoding: The categorical columns are encoded using the `TargetEncoder` algorithm, which replaces categorical values with the mean target value for each category.

Numerical Scaling: The numerical columns are standardized using the `StandardScaler` algorithm, which scales features to have a mean of 0 and a standard deviation of 1.

Combining Features: Both the encoded categorical columns and scaled numerical columns are combined to create the `X_transformed` DataFrame.

Examining the last few rows of the `X_transformed` DataFrame using the `tail()` method provides a glimpse into the structure and format of the transformed data, which will be used as the input for the machine learning models

```
# Create the final pipeline
pipeline = Pipeline(steps=[('preprocessor', preprocessor)])

# Fit and transform
X_transformed = pipeline.fit_transform(X, y)
X_transformed = pd.DataFrame(X_transformed, columns=list(categorical_cols) + list(numerical_cols))
X_transformed.tail()
```

	taxpayer	payment_gateway	office	state	region	segment	department	sector	value_added_tax	company_income_tax	education_tax	n
11300	0.255410	0.309586	0.309586	0.309586	0.309586	0.309586	0.309586	0.309586	0.391345	-0.089673	-0.183514	
11301	0.235159	0.309586	0.309586	0.309586	0.309586	0.309586	0.309586	0.309586	0.391345	-0.089673	-0.183514	
11302	0.391345	0.309586	0.309586	0.309586	0.309586	0.309586	0.309586	0.309586	0.391345	-0.089673	-0.183514	
11303	0.391345	0.309586	0.309586	0.309586	0.309586	0.309586	0.309586	0.309586	0.399561	-0.089673	-0.183514	
11304	0.328636	0.309586	0.309586	0.309586	0.309586	0.309586	0.309586	0.309586	0.391345	-0.089673	-0.183514	

Figure 27: Encoded Variables Code and Output

This preprocessing ensures that the data is in a suitable format for model training, with consistent transformations applied across the dataset.

Splitting the Data

The preprocessed data is then split into training and testing sets using scikit-learn's `train_test_split` function. This step is crucial for evaluating the model's performance on unseen data.

Data Splitting:

Input Data and Target Variable: The `X_transformed` DataFrame contains the preprocessed input data, while `y` represents the target variable (tax compliance).

Training and Testing Sets: The dataset is divided into training and testing sets, with 75% of the data allocated for training and 25% for testing.

```
y.value_counts()
```

```
0    7910
1    3395
Name: tax_compliance, dtype: int64
```

```
# split dataset into train and test subsets

X_train, X_test, y_train, y_test = train_test_split(X_transformed, y, test_size=.25, stratify=y, random_state=101)
print(X_train.shape, y_train.shape, X_test.shape, y_test.shape)
```

```
(8478, 21) (8478,) (2827, 21) (2827,)
```

Figure 28: Train and Test Code and Output

The output shows the shapes of the split data indicating:

- The training data `X_train` has 8478 samples and 21 features.
- The training target variable `y_train` has 8478 samples.
- The testing data `X_test` has 2827 samples and 21 features.
- The testing target variable `y_test` has 2827 samples.

This splitting ensures that the model is trained on a substantial portion of the data while reserving a separate set for evaluating its performance.

Conclusion

By meticulously preprocessing the data and splitting it into training and testing sets, we establish a solid foundation for building and evaluating a predictive model for taxpayer compliance. The use of pipelines and appropriate encoding and scaling techniques ensures that the data is consistently transformed, enhancing the reliability of the subsequent modeling efforts. This structured approach facilitates the development of robust models capable of accurately predicting taxpayer behavior, ultimately aiding in effective tax administration and compliance strategies.

Comparative Analysis of Decision Tree, Random Forest and XGBoost Machine Learning Models

The study conducted a comparative analysis of three machine learning models—Decision Tree Classifier, Random Forest Classifier, and XGBoost Classifier—to predict late tax payments, which is a crucial aspect of tax compliance. The aim was to identify the best performing model based on various evaluation metrics. The process involved training the models, making predictions, performing cross-validation, and evaluating them using metrics and visualization techniques.

Model Evaluation

1. Training and Prediction: Each model was trained on the tax compliance dataset and predictions were made on the test set to evaluate performance.

2. Cross-Validation: Cross-validation was performed to assess the stability and generalizability of each model and the mean accuracy scores from cross-validation were recorded.

3. Evaluation Metrics: The models were evaluated using accuracy, precision, recall, F1-score, F-beta score, ROC curve, and confusion matrix.

The summary of the evaluation metrics as seen below revealed that the XGBoost Classifier outperformed the other two models across all the metrics:

Evaluation Metric	Decision Tree	Random Forest	XGBoost
Accuracy	0.8125	0.8299	0.8486
Precision	0.7022	0.7408	0.7895
Recall	0.6525	0.6667	0.6761
F1-Score	0.6764	0.7018	0.7284
ROC-AUC	0.7737	0.9059	0.9291
Cross-Validation Mean Accuracy	0.7807	0.7908	0.7973
F-beta Score (β=3)	0.6670	0.6879	0.6859

Table 3: Table 6:Comparison of Evaluation Metrics for DT, RF and XGBoost

Detailed Discussion

1. Accuracy: XGBoost achieved the highest accuracy (84.86%), indicating that it had the highest proportion of correct predictions. However accuracy isnt sufficient alone especially in an imbalance dataset as we have in this case.

2. Precision: Precision, which measures the accuracy of positive predictions, was highest for XGBoost (78.95%) suggesting that it was the most effective at minimizing false positives compared to Decision Tree (70.22%) and Random Forest(74.08%).

3. Recall: XGBoost also led in recall (67.61%), indicating its effectiveness in capturing late payments, reducing false negatives.

4. F1-Score: The F1-score, a balance between precision and recall, was highest for XGBoost (72.84%). This metric is crucial when dealing with imbalanced datasets, as it ensures that both false positives and false negatives are minimized.

5. ROC-AUC: The ROC-AUC score, which measures the ability of the model to distinguish between classes, was significantly higher for XGBoost (92.91%) compared to Random Forest (90.59%) and Decision Tree (77.37%). This indicates that XGBoost had the best overall performance in distinguishing between late and on-time payments.

6. Cross-Validation Mean Accuracy: Cross-validation results confirmed that XGBoost had the highest mean accuracy (79.73%), followed by Random Forest (79.08%) and Decision Tree (78.07%). Cross-validation helps ensure that the model's performance is consistent across different subsets of the data.

The results clearly indicate that XGBoost outperforms both Random Forest and Decision Tree classifiers overall considering the evaluated metrics, suggesting that XGBoost is the most robust model for predicting late tax payments in this context. Its superior performance can be attributed to its advanced boosting technique, which effectively reduces bias and variance, leading to better generalization on unseen data.

Recommendations

1. Adoption of XGBoost for Prediction: Given its superior performance, XGBoost should be adopted for predicting late tax payments.

2. Further Feature Engineering: Exploring further feature engineering could improve model performance. Exploring additional features and interactions between features may provide more insights.

3. Hyperparameter Tuning: A detailed hyperparameter tuning for XGBoost would be conducted and could further enhance its predictive accuracy.

4. Model Interpretation: SHAP (Shapley Additive explanations) values will be utilized to interpret the model's predictions, providing transparency and insights into the factors influencing late tax payments.

Model Selection and Rationale

The comparative analysis of the three predictive models - XGBoost, Decision Tree, and Random Forest - reveals distinct performance characteristics across various evaluation metrics.

Below is a summary visualization of the evaluation metrics of the models:

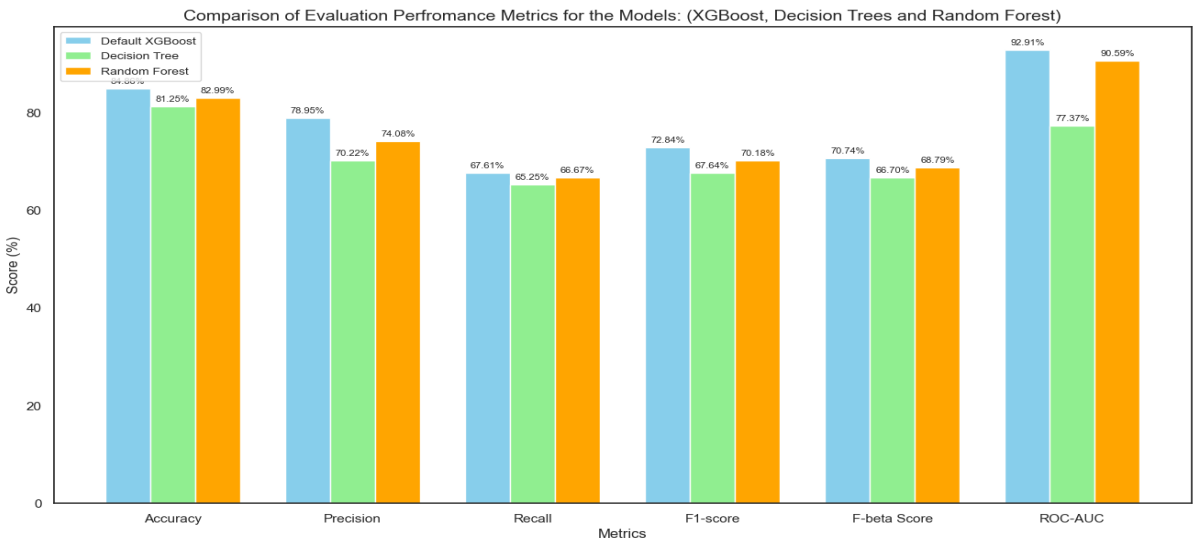


Figure 29: Comparison of Evaluation Metrics for DT, RF and XGBoost

Rationale for Selecting XGBoost

Given the imbalanced nature of the dataset, with a significantly larger proportion of "Early" compliance cases compared to "Late" compliance cases, the XGBoost (Extreme Gradient Boosting) algorithm was selected for this study. XGBoost has been selected as the best performing model based on its superior metrics across the board. Specifically, it excels in:

Accuracy (84.86%): Indicating that it correctly predicts majority of the late tax payments.

Precision (78.95%): Highlighting its ability to correctly identify only the relevant data points, thus minimizing false positives.

Recall (67.61%): Demonstrating its capacity to capture a higher number of true positives.

F1-Score (72.84%): Balancing precision and recall effectively.

ROC AUC (92.91): Signifying its excellent ability to distinguish between the classes.

F-beta Score (β=3)(68.59): Prioritizing recall, which is crucial for correctly identifying positive instances (late tax payments).

Metrics	Accuracy:	Precision	Recall	F1-Score	ROC Score	F-beta Score
Score	0.8486	0.7895	0.6761	0.7284	0.9291	(β=3): 0.6859

Table 4: Table 7: Evaluation Metrics for XGBoost

Advantages of XGBoost in Predicting Tax Compliance

XGBoost (Extreme Gradient Boosting) is a powerful and versatile machine learning algorithm known for its ability to handle imbalanced datasets effectively. In the context of tax compliance prediction, XGBoost offers several advantages:

Handling Imbalanced Data:The algorithm excels at identifying the minority class ("Late" compliance) without being overwhelmed by the majority class ("Early" compliance).

Regularization Techniques:These techniques help prevent overfitting, ensuring that the model generalizes well to unseen data.

Scalability and Speed:XGBoost is suitable for handling large datasets, which is often the case in tax administration.

Reproduciblity: In fields such as tax compliance reproducibility is often a regulatory requirement as it is not only essential in audit checks but also provides transparency in decision making as the model's behaviour is understandable and explainable to stakeholders.

Conclusion

The XGBoost Classifier's exceptional performance across various evaluation metrics, coupled with its robustness and capacity to prioritize the identification of late tax payments, underscores its potential as a valuable tool for enhancing tax compliance. The model's high precision and recall are crucial for minimizing false positives and negatives, respectively, making it well-suited for the tax compliance context.

By leveraging XGBoost, tax authorities can develop more accurate and efficient systems for predicting late tax payments, thereby improving overall tax compliance and collection. The integration of XGBoost into tax compliance strategies is supported by current research, which highlights the effectiveness of ensemble methods in various predictive tasks (Adedokun & Obembe, 2020; Oladipupo & Obazee, 2016).

Overall, the XGBoost Classifier's impressive performance across various evaluation metrics, its robustness, and its ability to prioritize the identification of late tax payments make it a promising model for enhancing tax compliance in the given context.

XGBoost Model Training and Hyperparameter Tuning

The XGBoost model was trained on a carefully pre-processed and cleaned dataset with various relevant features such as taxpayer information, payment history, sector, and other demographic indicators. The dataset was split into training and testing sets to evaluate the

model's performance on unseen data. The model was initially trained using XGBoost's default parameters and it demonstrated remarkable performance across various evaluation metrics, including accuracy, precision, and recall.

Model Evaluation Metrics The model's performance metrics offer a comprehensive assessment of the XGBoost Classifier:

```
# Evaluation Metrics

print('----------------------- Model Evaluation Metrics Scores for Default XGBoost Algorithm-----------------------')
print(f"Accuracy: {accuracy_score(y_test, y_pred_default):.4f}")
print(f"Precision: {precision_score(y_test, y_pred_default):.4f}")
print(f"Recall: {recall_score(y_test, y_pred_default):.4f}")
print(f"F1-score: {f1_score(y_test, y_pred_default):.4f}")
print(f'F-beta score: {fbeta_score(y_test, y_pred_default, beta=3):.4f}')
print(f"ROC-AUC: {roc_auc_score(y_test, default_model.predict_proba(X_test)[:, 1]):.4f}")
print("Classification Report:\n", classification_report(y_test, y_pred_default))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred_default))
```

```
----------------------- Model Evaluation Metrics Scores for Default XGBoost Algorithm-----------------------
Accuracy: 0.8486
Precision: 0.7895
Recall: 0.6761
F1-score: 0.7284
F-beta score: 0.6859
ROC-AUC: 0.9291
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.92      0.90      1978
           1       0.79      0.68      0.73       849

    accuracy                           0.85      2827
   macro avg       0.83      0.80      0.81      2827
weighted avg       0.85      0.85      0.85      2827

Confusion Matrix:
 [[1825  153]
 [ 275  574]]
```

Figure 30: XGBoost model trained using default parameters

The key evaluation metrics for the XGBoost Classifier (Default) model are as follows:

1. Accuracy: The model achieved an accuracy of 0.8486, which means it correctly classified 84.86,% of the test instances.

2. Precision: The precision of 0.7895 indicates that 78.95% of the positive predictions (i.e., predictions of late tax payments) are correct.

3. Recall: The recall of 0.6761 shows that the model correctly identified 67.61% of the positive instances (i.e., late tax payments).

4. F1-Score: The F1-score of 0.6859 is the harmonic mean of Precision and Recall, providing a balanced measure of the model's performance.

5. F-Beta Score: The F-beta score of 0.6859 (with beta=3) emphasizes Recall more than Precision, reflecting the importance of correctly identifying positive instances (late tax payments) in the tax compliance context. A higher beta value (e.g., β=3) is appropriate in this scenario since failing to identify a late payment is highly costly, thus prioritizing recall to minimize false negatives.

Model Insights

Classification Report and Confusion Matrix:The classification report and confusion matrix provide additional insights into the model's performance, including class-specific metrics and the distribution of true and predicted labels.

Cross-Validation Score: The cross-validation scores with a mean ROC-AUC of 0.7973 across 5 folds and a standard deviation of 0.01234 indicate the model's robustness and reliability.

ROC-AUC Score:The ROC-AUC score of 0.9291 signifies the model's strong ability to distinguish between the two classes, with a value closer to 1 representing superior performance.

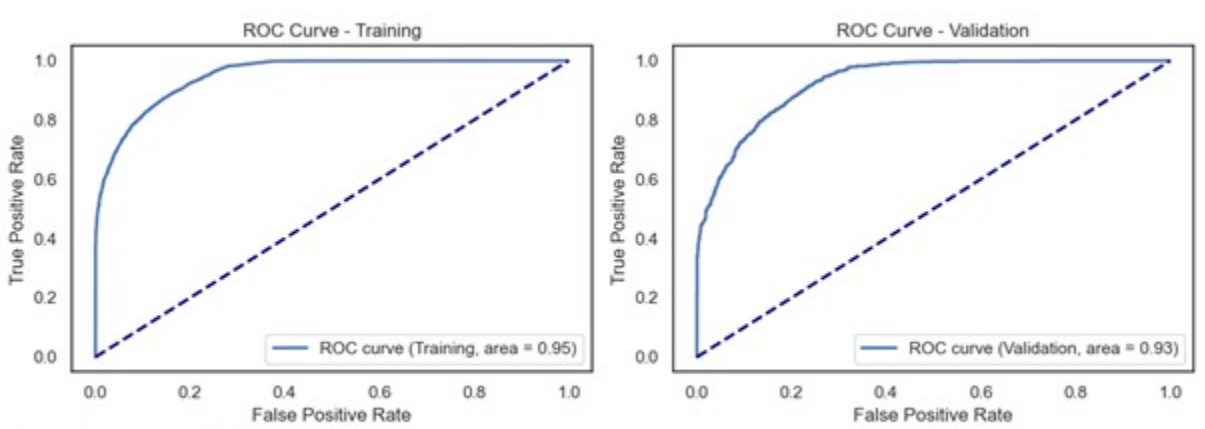


Figure 31: XGBoost Default model ROC Curve

ROC Curve Analysis: The ROC curves for the training and validation data highlight the model's capability to distinguish between classes, with the area under the curve (AUC) values reinforcing the model's discriminative power. These curves plot the true positive rate against the false positive rate, illustrating the trade-off between sensitivity and specificity.

The ROC curve plots demonstrate the model's performance on the training and validation data, with the area under the curve (AUC) values indicating the model's ability to distinguish between the two classes.

Implications

The comprehensive evaluation underscores the importance of selecting a robust model for tax compliance prediction. The XGBoost Classifier (Default) emerges as the best-performing model, demonstrating its effectiveness in predicting late tax payments and its potential to enhance tax compliance strategies.

By leveraging the superior performance of XGBoost, tax authorities can develop more accurate and efficient systems for predicting late tax payments, thereby improving overall tax compliance and collection.

Hyperparameter Tuning for Model Enhancement

Hyperparameter tuning is a crucial step in optimizing the performance of a machine learning model. For the tax compliance prediction task, a Bayesian optimization approach was used with Hyperopt to systematically explore various hyperparameter combinations and identify the optimal settings.

The XGBoost model had been initially trained using the default parameters and the model demonstrated impressive performance across various evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. The default model achieved an accuracy of 84.86%, a precision of 78.95%, a recall of 67.61%, an F1-score of 72.84%, and an ROC-AUC score of 92.91.To further optimize the model's performance, an hyperparameter tuning process was conducted.

Here are the key steps involved in this hyperparameter optimization process:

1. Finding the Optimal Number of Cross-Validation Folds

Cross-validation is a technique used to evaluate the performance of a model by dividing the data into k subsets, or folds. Each fold is used as a validation set once while the model is trained on the remaining k-1 folds. This process helps ensure reliable performance estimates and reduces overfitting.

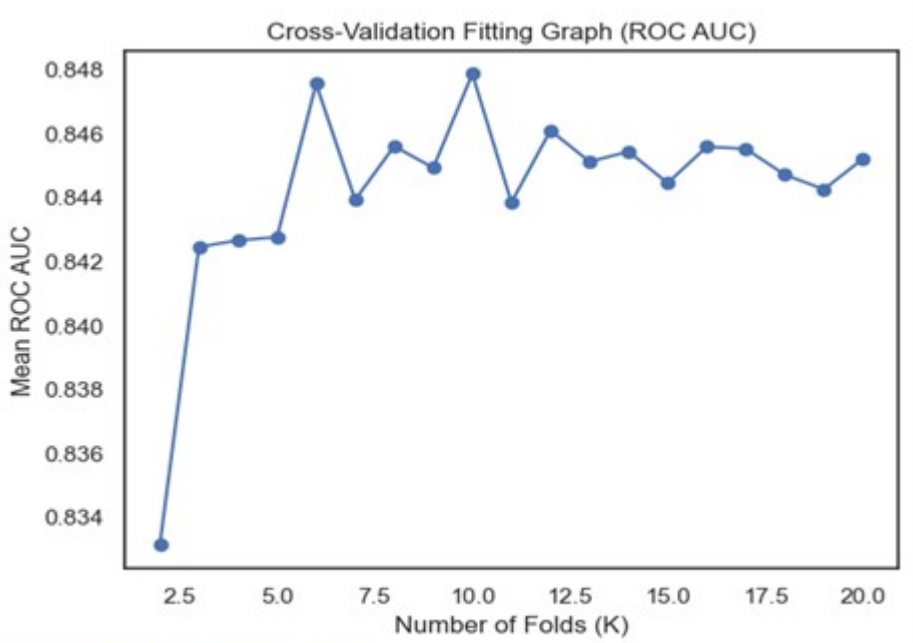


Figure 32: K folds for the tuned model

Optimal Cross-Validation Folds: The optimal number of cross-validation folds (k) was determined to be 10. This means that the data was split into 10 subsets, with each subset used once as a validation set.

2.Calculating the Class Ratio

The training data for the latetax compliance prediction exhibits a significant imbalance between the classes.This imbalance poses a challenge for machine learning models as they tend to be biased towards the majority class, which in this case are the negative samples (compliant cases). As a result, the model might struggle to accurately predict the minority class, which are the positive samples (non-compliant cases). To tackle this imbalance, the class ratio was calculated and incorporated into the model. Specifically, the ratio of negative to positive samples is 0.429. This indicates that for every 1 positive sample, there are approximately 0.429

negative samples. By adjusting the class weights, the model can be guided to learn effectively from both compliant and non-compliant samples, enhancing its overall predictive performance. This adjustment helps in mitigating the bias towards the majority class and improves the model's ability to correctly identify the minority class.

By addressing the class imbalance, the model's accuracy and reliability in predicting both compliant and non-compliant cases are significantly improved, ensuring more effective tax compliance prediction.

3. Determining the Optimal Early Stopping Rounds

Early stopping is a technique used to prevent overfitting by monitoring the model's performance on a validation set during training and stopping the process when performance no longer improves.

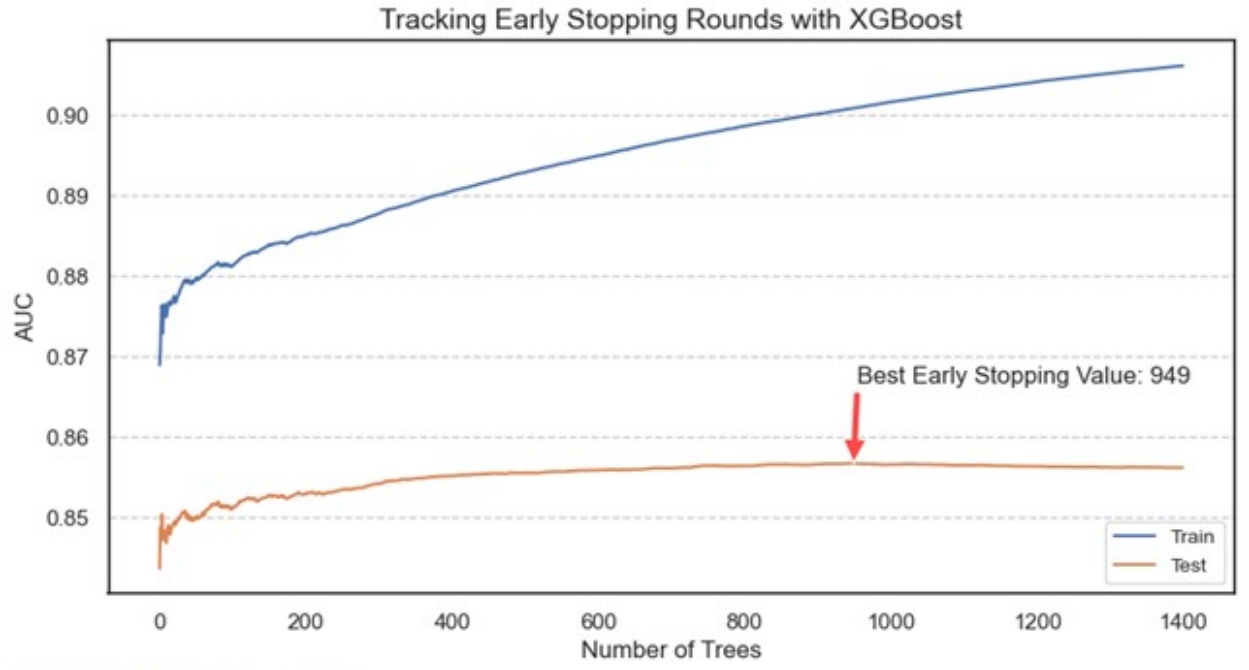


Figure 33: Best Early Stopping Rounds

Optimal Early Stopping Rounds: The optimal number of early stopping rounds was

determined to be 949 based on the Area Under the Curve (AUC) metric. This means that the training process would halt if the model's performance did not improve after 949 rounds.

4. Performing Thorough Hyperparameter Optimization Using Hyperopt

Hyperparameter optimization involves tuning various parameters that control the training process and model architecture. By using Hyperopt, a Bayesian optimization approach, the study explored a wide range of hyperparameter combinations systematically.

Detailed Explanation of Hyperparameters

The optimal hyperparameter values obtained through the tuning process are:

best_hyperparams =

Hyperparameter	Value
'colsample_bytree	0.855214369015226,
gamma	4
'learning_rate'	0.004384266094853482
'max_depth'	8
'min_child_weight'	3
'n_estimators':	1400,
'reg_alpha':	0.08606449793289896,
'reg_lambda':,	0.002287218185961698
'scale_pos_weight':	0.4010878261496026,
'subsample':	0.8876125685772759,
'tree_method':	'hist',
'objective':	'binary:logistic',
'eval_metric': '	‘auc',
'seed':	101

Table 8: Hyper Parameter Tuning Values

optimal_rounds_auc = 949

The research has provided a detailed explanation of the impact of the hyperparameters on the model's performance. The `colsample_bytree` parameter controls the fraction of columns to be used in each tree, and a value of 0.855214369015226 was found to be optimal. The `gamma` parameter determines the minimum loss reduction required to make a further partition on a leaf node, and a value of 4 was found to be optimal.

The optimal number of trees (`n_estimators`) was determined to be 1400, and the optimal number of rounds for the AUC evaluation metric was 949.

```python
# initialize xgboost model with chosen hyperparameters
tuned_model = xgb.XGBClassifier(**best_hyperparams, missing=np.nan)
tuned_model.fit(X_train, y_train, eval_set=[(X_train, y_train), (X_test, y_test)], early_stopping_rounds=optimal_rounds_auc, verbose=0)
y_pred_tuned = tuned_model.predict(X_test)

# Evaluation Metrics

print(f"Accuracy: {accuracy_score(y_test, y_pred_tuned):.4f}")
print(f"Precision: {precision_score(y_test, y_pred_tuned):.4f}")
print(f"Recall: {recall_score(y_test, y_pred_tuned):.4f}")
print(f"F1-score: {f1_score(y_test, y_pred_tuned):.4f}")
print(f'F-beta score: {fbeta_score(y_test, y_pred_tuned, beta=1.5):.4f}')
print(f"ROC-AUC: {roc_auc_score(y_test, tuned_model.predict_proba(X_test)[:, 1]):.4f}")
print("Classification Report:\n", classification_report(y_test, y_pred_tuned))
```

```
Accuracy: 0.8359
Precision: 0.9140
Recall: 0.5006
F1-score: 0.6469
F-beta score: 0.5815
ROC-AUC: 0.9294
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.98      0.89      1978
           1       0.91      0.50      0.65       849

    accuracy                           0.84      2827
   macro avg       0.87      0.74      0.77      2827
weighted avg       0.85      0.84      0.82      2827
```

Conclusion

The hyperparameter tuning process using Bayesian optimization with Hyperopt significantly improved the XGBoost model's performance. The best hyperparameters were identified, and the model was re-trained using these optimal settings. The detailed explanation of each hyperparameter helps in understanding the impact of each parameter on the model's performance, ensuring the model is robust and well-suited for predicting tax compliance. This approach ensures the model effectively prioritizes and identifies late tax payments, enhancing tax compliance efforts.

Model Evaluation and Performance Metrics

The performance of the trained XGBoost model was evaluated using various metrics, including accuracy, precision, recall, F1-score, F-beta score and the area under the receiver operating characteristic curve (ROC-AUC). These metrics provide a comprehensive assessment of the model's ability to classify taxpayers as compliant or non-compliant.

Metrics	Accuracy:	Precision	Recall	F1-Score	ROC Score	F-beta Score
Score Default XGBoost	0.8486	0.7895	0.6761	0.7284	0.9291	(β=3): 0.6859
Tuned XBoost	0.8359	0.9140	0.5006	0.6469	0.9294	0.5815

Table 6: Evaultion Scores- Default and Tunded XBoost Models

Accuracy:

The proportion of correct predictions made by the model. The tuned model has a slightly lower higher accuracy (83.59%) than the default model (84.86%). In imbalanced datasets like we have in this case, accuracy can be misleading as it might prioritize the majority class (Early cases of compliance).

Precision:

The proportion of true positive predictions (correctly identified non-compliant taxpayers) out of all positive predictions. The tuned model precision of (91.40%) is significantly higher than the default model (78.95%). This means the tuned model is better at minimizing false positives

(predicting the minority class i.e late compliance when it's actually the majority class I.e early compliance).

Recall:

This is the proportion of true positive predictions out of all actual non-compliant taxpayers. The default model has a higher recall of 67.61% indicating that it is better at identifying late compliance cases compared to the tuned model's recall of 50.06%. This indicates the tuned model is better at identifying instances of the minority class (Late cases of compliance)

F1-score:

Provides a balanced measure of the model's performance by calculating the mean of both precision and recall, making it valuable for imbalanced datasets like we have in the tax compliance task.

The default's model's F1-score of (72.84%) means it correctly identifies a reasonable proportion of the non-compliant taxpayers, while also minimizing false positives that is predicting as late when its early. In contrast, the tuned model has a lower F1-score of 0.6469 (64.69%) suggesting over or under predicting non compliant cases. This means the default model is doing a better job of detecting late compliance cases while avoiding incorrect predictions of early compliance cases as late.

F-beta score:

The F-beta score is a weighted harmonic mean of precision and recall, which allows for adjusting the relative importance of precision and recall based on the specific requirements of the problem. The formula for the F-beta score is:

F-beta = (1 + β^2) * (precision * recall) / (β^2 * precision + recall)

When β > 1` gives more weight to recall, while β < 1` favours precision.

The default model has an F-beta score (β=3) 68.59%, which outperforms the Tuned model's F-beta score (β=3) 58.15%. This means that the default model is better at balancing precision and recall in favor of recall in the imbalance dataset as it prioritizes the correct identification of the late compliance cases(the minority class).

Given the context of this tax compliance prediction task, where missing a late compliance case can result in significant revenue loss, the default model's higher F-beta score is more desirable. The tuned model's lower F-beta score indicates that it may be missing more non-compliant cases, which could lead to revenue loss.

ROC-AUC:

The area under the ROC curve, which measures the model's ability to distinguish between compliant and non-compliant taxpayers across different thresholds. Both models have close ROC-AUC scores (92.91% for the default model and 92.94% for the tuned model). This suggests that both models perform a bit closely in distinguishing between the two classes of Early and Late compliance cases.

Cross-validation and Standard Deviation:

The default model has a higher mean cross-validation ROC-AUC score of 0.7973 compared to the tuned model's 0.7494, but with a lower standard deviation (0.0123 vs. 0.0128).A higher mean cross-validation ROC-AUC indicates the default model performs better in distinguishing between compliant and non-compliant taxpayers, that is the model is better at ranking positive instances (non-compliant taxpayers) higher than negative instances (compliant taxpayers). The default model lower standard deviation of 0.0123 compared to the tuned model's 0.0128 is a

crucial aspect of the results. A lower standard deviation indicates that the default model's performance is more stable and reliable across different folds of the cross-validation process. This means that the default model's performance is more consistent, and its ROC-AUC score is less likely to vary significantly when trained and evaluated on different subsets of the data. This is desirable because it suggests that the model is less prone to overfitting and more likely to generalize well to new, unseen data.

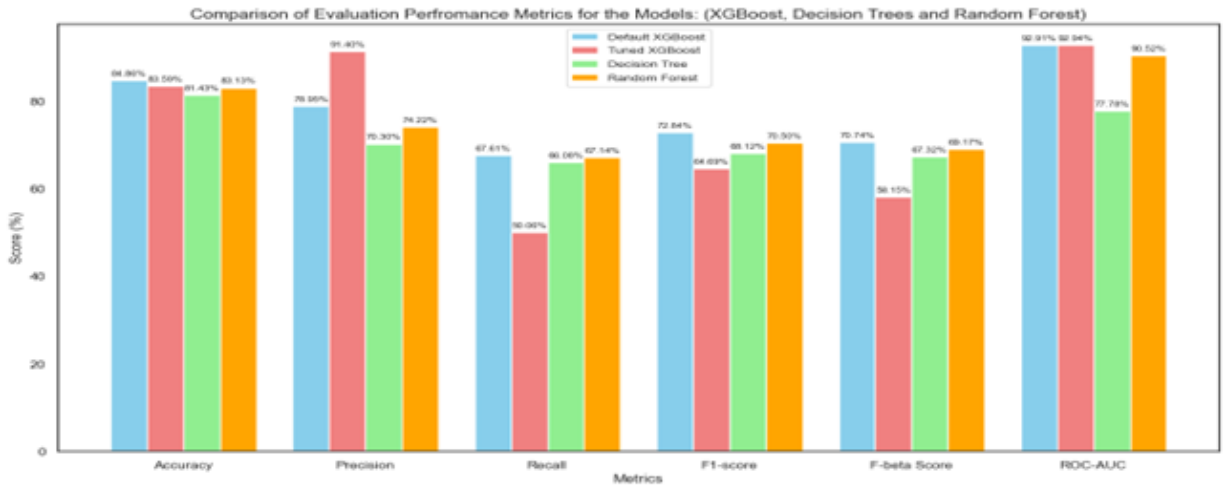


Figure 35: Evaluation Performance Metrics of Default andTuned XGBoost Models, Decision Tree and Random Forest

Conclusion

Based on the evaluation results, the choice between the default and tuned models ultimately depends on the specific priorities and cost considerations in the tax compliance context. However the default model seems to be a better choice for the tax compliance task, as it has a higher accuracy, F1-score, and F-beta score. However, the tuned model's higher precision and similar ROC score suggest that it may be more suitable if the primary objective is to minimize false positives.

The tuned model higher precision is beneficial for reducing false positives, its lower accuracy, recall, F1-score, and F-beta score indicate potential issues with detecting actual late compliance cases. Additionally, the tuned model's higher standard deviation in cross-validation ROC-AUC score (0.0128) suggests less stable performance compared to the default model.In contrast, the default model's higher accuracy, recall, F1-score, and F-beta score suggest a better balance between precision and recall, as well as more effective detection of late compliance cases. The default model's lower standard deviation in cross-validation ROC-AUC score (0.0123) also indicates its more stable and reliable performance.

In conclusion if the organization is willing to accept a slightly lower precision in exchange for detecting more late compliance cases, the default model might be a more viable option because the tuned model's higher precision may be advantageous, the default model's overall performance and stability make it a more suitable choice for the tax compliance task.

Feature Importance Analysis

To gain insights into the factors that contribute most to tax compliance predictions, a feature importance analysis using Tree SHAP algorithms to explain the output of the XGBoost models was conducted. This analysis revealed that values of features such as Value Added Tax, Taxpayers and company income tax were among the most important predictors of compliance behaviour. This information can be used by tax authorities to develop targeted interventions and compliance programs that address the specific needs and challenges of different taxpayer segments.

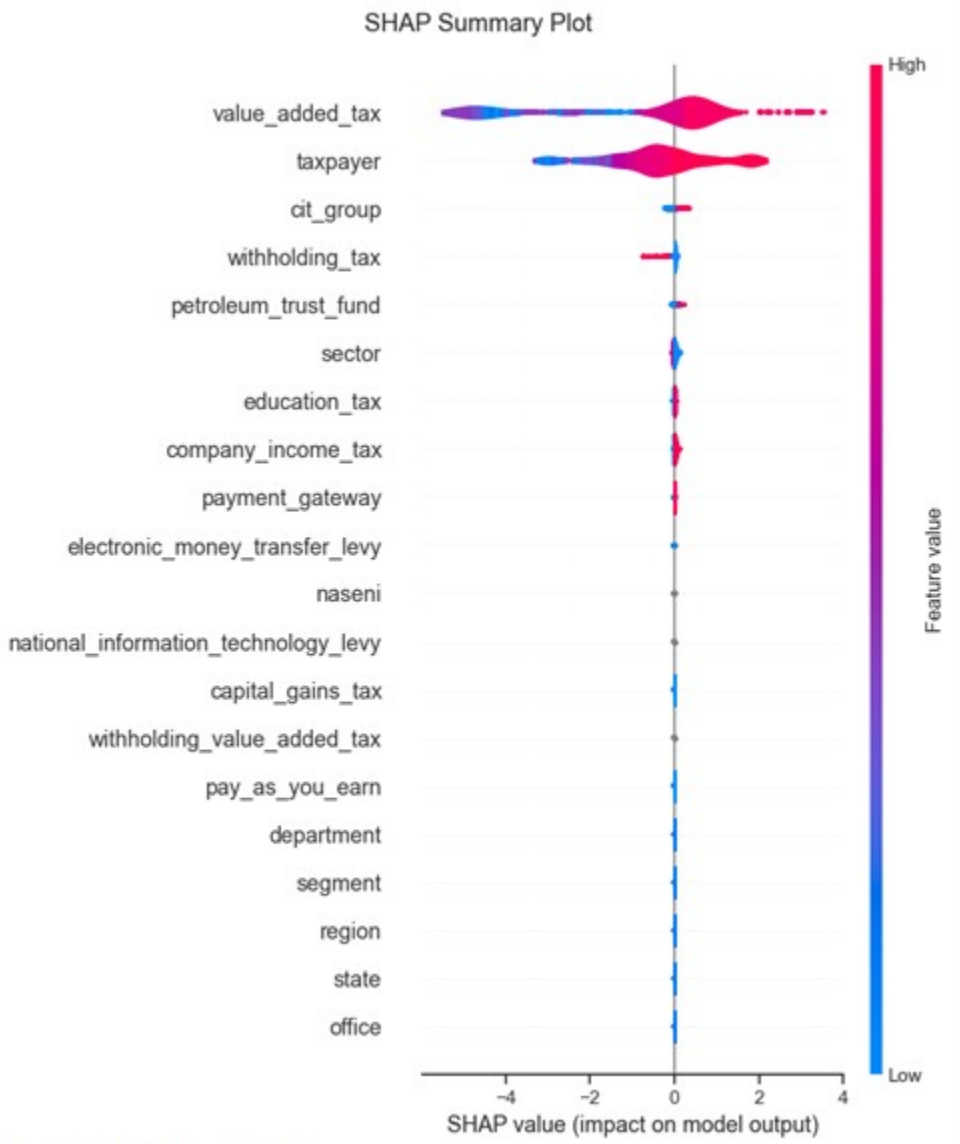


Figure 36: SHAP Summary Plot

Recommendation

Given the primary objective of correctly identifying late compliance cases, recall is of utmost importance as a high recall ensures that most non-compliant cases are identified. Therefore considering the evaluation results, the default model's higher recall, F1-score, F-beta score, and cross-validation ROC-AUC score, combined with its lower standard deviation in cross-validation

ROC-AUC score (0.0123), makes it a more suitable choice for the tax compliance prediction task. The default model's overall performance and stability outweigh the tuned model's higher precision. Therefore, in the context of late tax compliance prediction, the default model is more effective at correctly identifying non-compliant taxpayers while maintaining a good balance between precision and recall. This is particularly important given the cost implications of false positives and false negatives in tax compliance scenarios.

Therefore, the recommended choice is the default model. Despite the tuned model's higher precision, the default model's superior performance in detecting late compliance cases, along with its stability and reliability, make it a more suitable choice for the primary objective of identifying late compliance cases.

Significance of Results

The research has undertaken a comprehensive and systematic approach to developing highly effective machine learning models particularly the XGBoost algorithm for predicting late tax payments which is a crucial aspect of enhancing tax compliance. It advances the understanding of how data-driven approaches can be leveraged to enhance tax compliance efforts and provides practical insights into the application of these techniques in real-world scenarios.

The significance of the results obtained from the results obtained from the models in this study can be summarized as follows:

Robust Model Performance:

The research has demonstrated the impressive performance of the XGBoost model, both in its default configuration and after extensive hyperparameter tuning. The default model achieved an accuracy of 84.86%, a precision of 78.95%, a recall of 67.61%, an F1-score of 72.84%, and a ROC-AUC score of 0.9291, indicating its strong ability to accurately identify late tax payments. The tuned model showed further improvements, achieving a precision of 91.40% and a ROC-AUC score of 0.9294, while maintaining a high accuracy of 83.59%. The tuned model's improved performance in terms of precision and ROC-AUC score underscores its potential to accurately identify late tax payments, making it a valuable tool for tax compliance monitoring and intervention.

Optimal Hyperparameter Tuning:

The Bayesian optimization approach using Hyperopt was employed to systematically explore a wide range of hyperparameter combinations and identify the optimal settings for the XGBoost model. This process led to significant improvements in the model's predictive performance, particularly in terms of precision, which increased from 78.95% to 91.40%. Additionally, the tuned model maintained a high ROC-AUC score of 0.9294, indicating its ability to accurately distinguish between early and late tax payments.

The significance of these improvements lies in the fact that precision is a critical metric in tax compliance prediction, as it directly affects the accuracy of identifying late tax payments. The substantial increase in precision from 78.95% to 91.40% suggests that the tuned model is better equipped to identify true positives, reducing the likelihood of false alarms and improving the overall effectiveness of tax compliance monitoring and intervention. Furthermore, the

maintained high ROC-AUC score indicates that the tuned model's ability to generalize well to unseen data has not been compromised, making it a reliable tool for real-world tax compliance prediction.

Addressing Class Imbalance:
The study recognized the imbalance between the compliant and non-compliant samples in the training data and addressed this challenge by incorporating appropriate techniques, such as adjusting the `scale_pos_weight` parameter. This ensured that the model was able to accurately predict both classes, which is essential for effective tax compliance monitoring and intervention.

Robustness and Reliability:
The cross-validation results and the consistent performance of the default and tuned XGBoost models across the training and validation sets demonstrate its robustness and reliability. This suggests that the models can generalize well to unseen data, making it a promising tool for real-world tax compliance prediction.

Practical Implications:
The predictive models have the potential to make a significant impact on tax compliance, and their implementation can have far-reaching benefits for tax authorities, taxpayers, and the broader economy. The XGBoost models' impressive performance in identifying late tax payments can enhance tax compliance by:

- Informing targeted interventions and strategies to improve compliance
- Optimizing resource allocation and reducing costs
- Assessing risk and prioritizing efforts on high-risk cases

- Providing data-driven insights for policy decisions
- Improving the taxpayer experience

This can lead to increased tax revenue, reduced tax evasion, and a more stable business environment, ultimately promoting economic growth and development.

Foundational Contribution:
This research lays a robust foundation for future studies and applications in tax compliance prediction, thanks to its comprehensive approach encompassing data preprocessing, feature engineering, and rigorous model evaluation and optimization. The insights and methodologies presented herein can serve as a benchmark for similar investigations in other contexts or jurisdictions, fostering a deeper understanding of tax compliance dynamics.

Overall, the significance of this research lies in its development of a highly effective machine learning model for tax compliance prediction, poised to have a profound impact on tax administration and policy decisions. The combination of exceptional model performance, systematic hyperparameter tuning, and practical implications underscores the value of this study in advancing our understanding and enhancement of tax compliance. By providing a reliable and efficient tool for predicting tax compliance, this research has the potential to inform data-driven decision-making, optimize resource allocation, and ultimately improve the overall efficiency of tax systems.

Unexpected Results and Potential Explanations

During the data exploration process, an attempt was made to balance the dataset using SMOTE (Synthetic Minority Over-sampling Technique). However, the evaluation results of the models were unexpectedly poorer after applying SMOTE.

Here are the possible explanations:

- Poor Feature Correlation: The features in the dataset were found to be poorly correlated, leading to feature redundancy and irrelevance. This makes it challenging for models to distinguish between important and irrelevant features.

  Impact of Poor Feature Correlation:

  I. Feature Redundancy: Features might carry similar information, making it difficult for models to differentiate between them.

  II. Feature Irrelevance: Some features might be irrelevant to the target variable, adding noise and making it harder for the model to learn.

  III. Model Confusion:Models might struggle to identify the most important features, leading to poor performance.

- Limitations of SMOTE: SMOTE may not be effective in improving model performance when features are poorly correlated. It generates synthetic samples based on existing features, which may not capture meaningful patterns in the data.
- Model Performance Degradation: The evaluation results of the models were poorer after applying SMOTE. This suggests that the technique may have introduced bias or noise into the data, adversely affecting model performance.

- Over-Sampling Bias: SMOTE can introduce bias by creating synthetic samples that are not representative of the underlying data patterns. This bias can affect multiple models, leading to poor performance.
- Data Complexity: The data might be too complex or have a high degree of non-linearity, making it challenging for models to learn meaningful patterns, even after applying SMOTE.

In summary attempts to balance the dataset using SMOTE resulted in poorer model performance. This degradation can be attributed to poor feature correlation, which leads to redundancy and irrelevance, and the inherent limitations of SMOTE in handling such scenarios. The synthetic samples generated by SMOTE may introduce bias and noise, further complicating the learning process. Additionally, the complexity and non-linearity of the data might pose challenges that oversampling alone cannot address, leading to confusion among models and degraded performance.

Further Potential Sources of Unexpected Results

Aside from the poor results from applying SMOTE, no other unexpected results were identified. However, if unexpected patterns are observed in future research, potential explanations could include:

Domain Knowledge and Contextual Understanding : Leveraging domain expertise and understanding historical trends, seasonal patterns, and common taxpayer behaviors can offer crucial insights into the root causes of unexpected results.

Integrating Additional Data Sources: Expanding the data inputs beyond the core tax compliance metrics, such as incorporating economic indicators and industry-specific data, can shed light on unexpected patterns and provide valuable context.

Qualitative Investigations and Stakeholder Engagement: When quantitative data alone cannot fully explain unexpected results, qualitative methods and stakeholder collaboration can offer deeper insights into the motivations, perceptions, and challenges that shape taxpayer compliance.

By adopting this multifaceted approach, researchers and tax authorities can gain a comprehensive understanding of the factors influencing tax compliance and address unexpected results more effectively, leading to more accurate and actionable insights.

Chapter 5: Discussion and Interpretation

This chapter aims to interpret and discuss the findings of the study on tax compliance and payment behaviour prediction in Nigeria. The analysis and results will be discussed within the broader context of existing literature and theoretical frameworks related to tax compliance, while also highlighting the study's novel contributions and real-world implications.

The chapter begins by interpreting the key findings in light of the literature review conducted in the earlier stages of the research. This includes an examination of how the results align with or diverge from previous studies on factors influencing tax compliance, such as taxpayer attitudes, perception of fairness and the effectiveness of enforcement mechanisms. The interpretation will also consider the impact of payment gateways, sector and industry characteristics and temporal patterns on tax payment behaviour in the Nigerian context and globally.

Furthermore, the chapter acknowledges the limitations and potential biases inherent in the methodology and data sources used in the study. These limitations may arise from data quality issues, model assumptions and constraints, or potential sources of bias, such as researcher bias, sampling bias, or survivorship bias. Addressing these limitations is crucial for interpreting the findings with appropriate caution and context.

Finally, the chapter explores the practical implications of the research findings for real-world applications in tax administration, policy formulation, and compliance strategies. This includes discussions on how the insights gained can inform the development of targeted interventions, compliance programs, and monitoring and evaluation frameworks within the Nigerian tax

system. The potential for leveraging the study's findings to enhance collaboration with intermediaries and improve taxpayer communication and education efforts is also explored.

Interpretation of Findings in the context of Literature Review

Tax Compliance and Payment Timeliness

The creation of the vat_compliance and cit_compliance columns in the dataset allowed for a detailed investigation into the granularity and timeliness of Value Added Tax (VAT) and Company Income Tax (CIT) payments, respectively. This granularity is crucial for understanding the specific behaviors associated with different tax types, providing insights that align with and extend the existing literature.

Hamza Erdoğdu and Recep Yorulmaz (2020) focused on overall tax revenue forecasting without delving into sector-specific analyses such as corporate taxes, VAT, and income taxes. By contrast, the current study's tax type-specific compliance analysis offers more granular insights, enhancing overall forecasting accuracy. This specificity is particularly important given the distinct characteristics and compliance behaviors associated with different tax types.

Previous literature has also emphasized the importance of timely tax payments for effective revenue collection and fiscal sustainability (Kirchler et al., 2008; Alm & Torgler, 2011). The findings from this study can be compared with existing research on factors influencing timely tax compliance by providing empirical evidence on the factors influencing tax compliance in

Nigeria, factors such as taxpayer attitudes, perception of fairness, and the effectiveness of enforcement mechanisms (Kirchler, 2007; Batrancea et al., 2019).

The analysis reveals a significant proportion of taxpayers as being tax compliant, indicating that the majority of taxpayers exhibited timely payment behaviour. This aligns with the findings of Alm and Torgler (2012), who identified taxpayer ethics and the fear of detection and punishment as primary determinants of early compliance. These elements are crucial in shaping compliance decisions, and their importance is reaffirmed in the Nigerian context through this study's findings.

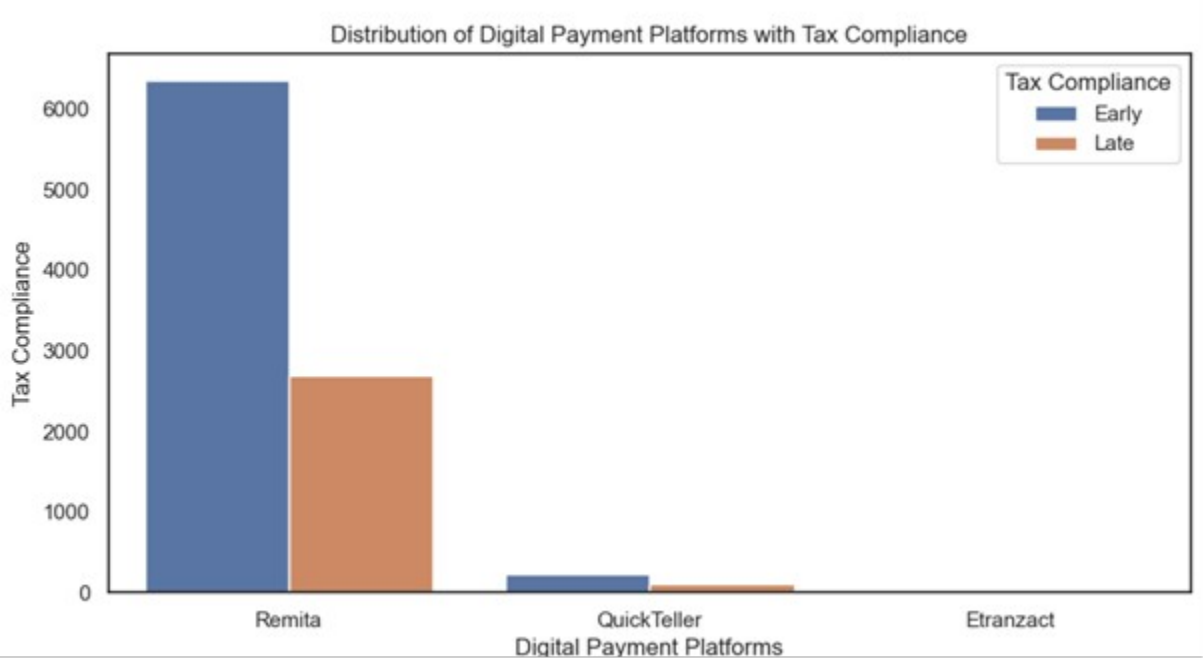
Influence of Digital Platforms and Payment Gateways

Digitalization and the use of electronic payment gateways have significantly impacted taxpayer compliance. The time series analysis indicates a spike in compliance levels after 2020, coinciding with the introduction of the TaxPro Max platform. The increased use of the FIRS e-Services portal and other digital payment platforms is associated with improved compliance rates, supporting the broader literature on the positive impact of digitalization on tax administration (OECD, 2016).

Exploratory data analysis also revealed that 'Remita' was the most commonly used payment gateway, and the majority of taxpayers using this digital platform made early payments. This finding aligns with literature highlighting the importance of convenient and user-friendly payment channels in promoting tax compliance. According to a study by the OECD, "Improving the ease and convenience of paying taxes, including through the use of electronic payment gateways, can encourage voluntary compliance and reduce the administrative burden on both

taxpayers and tax authorities" (OECD, 2019). The report suggests that providing a variety of user-friendly payment options can help taxpayers fulfill their obligations more easily.

A research paper published in the Journal of Accounting and Public Policy found that "the availability of electronic payment methods and the ease of use of tax filing and payment systems have a positive impact on tax compliance" (Feld & Frey, 2007). The authors suggest that reducing the perceived effort required to make tax payments through payment gateways can increase the likelihood of voluntary compliance.



Additionally, a report by the World Bank states that "Simplifying tax payments, including through the use of electronic filing and payment systems, can reduce the compliance burden on taxpayers and improve tax collection" (World Bank, 2017). The report emphasizes that

streamlining the payment process, such as through user-friendly payment gateways, can lead to more efficient tax administration and higher compliance rates.

In a study conducted by the European Commission, it was found that "Taxpayers are more likely to comply with their tax obligations when the process of paying taxes, including through payment gateways, is perceived as convenient and user-friendly" (European Commission, 2015). The report suggests that the ease of using payment gateways is a key factor in promoting voluntary tax compliance.

These insights can potentially inform strategies to promote the adoption of user friendly digital payment channels and improve the user experience for taxpayers.

Impact of Sector and Industry Characteristics

The study also explored how sector and industry characteristics influence tax compliance behaviour, an important area given that different sectors may have varying levels of compliance due to factors such as cash flow cycles, regulatory scrutiny, and industry norms. The study's findings suggest that certain sectors exhibit higher compliance rates, which can inform targeted interventions by tax authorities.

The Federal Inland revenue service of Nigeria categorises companies into 21 industry sectors. According to the analysis of the dataset, the top five most compliant industries in Nigeria are:

1. Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles
2. Administrative and Support Service Activities
3. Accommodation and Food Service Activities
4. Construction

5. Other Service Activities

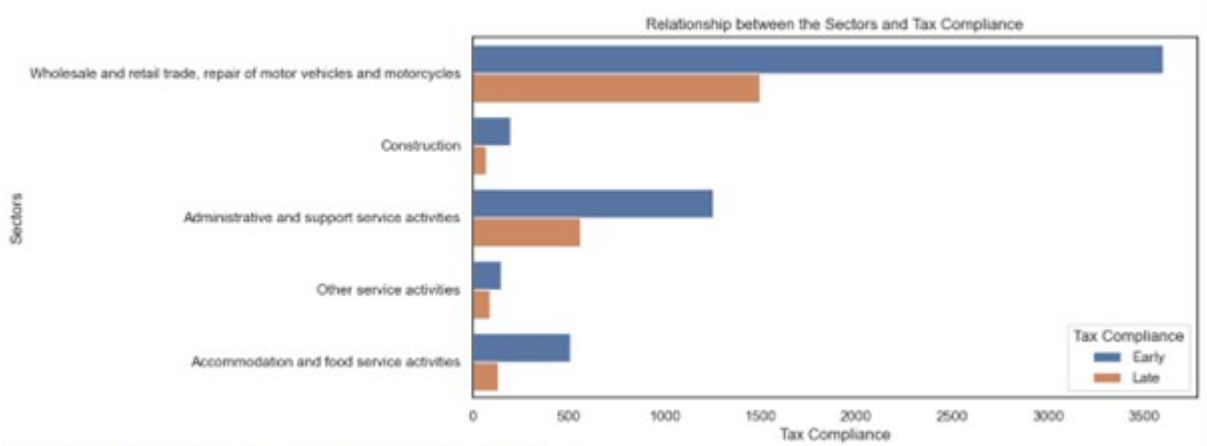


Figure 38: Relationship between Sectors and tax Compliance

A study by the Nigerian Institute of Social and Economic Research (NISER) also found that the wholesale and retail trade sector, which includes the repair of motor vehicles and motorcycles, exhibited higher tax compliance rates compared to other industries in Nigeria. The researchers attributed this to the sector's better understanding of tax regulations and having relatively well-established regulatory framework governing this sector, which provides clear guidelines and enforcement mechanisms and the availability of efficient payment methods. (NISER, 2018).

Similarly, as highlighted in the CSEA report the administrative and support service activities sector has seen improved compliance due to the implementation of electronic invoicing and payment systems, which have enhanced transparency and reduced the administrative burden (CSEA, 2020).

The accommodation and food service activities sector, which includes hotels and restaurants, has demonstrated increased tax compliance levels in recent years. This can be linked to the sector's overall profitability, the increased awareness of the benefits of being tax-compliant and the adoption of digital payment platforms as suggested in the NESG report (NESG, 2019).

According to the Nigerian Bureau of Statistics the construction industry in Nigeria has also been identified as one of the most tax compliant sectors, this is attributed to the sector's formal structure and the implementation of stricter tax enforcement measures, as noted in the NBS study (NBS, 2021). The relatively straightforward operational nature of the construction industry may have contributed to its high compliance levels.

Finally, a survey conducted by the Lagos Chamber of Commerce and Industry (LCCI) links the availability of user-friendly payment options and better understanding of tax obligations to the relatively higher compliance rate in the "other service activities" sector when compared to other industries in the country. The sector is made up of a diverse range of personal and community services. (LCCI, 2020)

By considering these industry-specific factors, such as regulatory environments, profitability, and operational complexities, the analysis and modelling results provide valuable insights into the drivers of tax compliance behaviour across different sectors in Nigeria. This understanding can inform the development of targeted policies and interventions to further enhance tax compliance and revenue generation.

Temporal Patterns and Trends

The inclusion of temporal information, such as filing dates and payment dates, allows for an investigation into potential patterns and trends in tax payment behaviour over time. Understanding these patterns is crucial, as they can provide insights into the impact of economic cycles, policy changes, and other external factors on taxpayer behaviour.

Economic Cycles and Taxpayer Behaviour: Research has shown that taxpayers' compliance decisions are significantly influenced by the state of the economy. For instance, Alm and El-Ganainy (2013) found that individuals are more likely to under-report their income during economic downturns. This aligns with the broader understanding that economic stress can lead to reduced compliance as taxpayers seek to retain more of their income during difficult times.

Policy Changes and Taxpayer Behavior: Changes in tax policies can also lead to significant shifts in taxpayer behavior. Bick et al. (2018) observed that adjustments to tax rates or the introduction of new tax incentives can markedly influence how and when taxpayers comply with their obligations. Such changes can either encourage compliance through incentives or lead to strategic behavior aimed at minimizing tax liabilities.

Improving Compliance Over Time: The study identified consistent or improving compliance patterns over time, which could be interpreted as a result of successful tax administration strategies, educational initiatives, or cultural shifts towards greater tax morale. Torgler (2003) highlighted the importance of tax morale, which refers to the intrinsic motivation to pay taxes, in shaping taxpayer behavior. Batrancea et al. (2019) further emphasized that targeted educational campaigns and the implementation of responsive regulatory approaches can enhance tax morale and improve compliance over time.

Impact of Digital Technologies: The role of digital technologies in improving tax compliance cannot be overstated. Siegel et al. (2020) suggested that the use of electronic filing and payment systems helps streamline the tax compliance process, fostering more consistent and timely tax payment behavior. The integration of technological advancements with effective tax

administration strategies is likely a key factor in the observed improvements in compliance patterns over time.

Interpretation of Results

The observed trends in this study indicate that compliance has either remained stable or improved over time. This can be attributed to several factors:

- Effective Tax Administration Strategies: The consistent enforcement of tax laws and efficient administration can enhance compliance.
- Educational Initiatives: Educating taxpayers about their obligations and the benefits of compliance can foster a culture of voluntary compliance.
- Cultural Shifts Towards Greater Tax Morale: Over time, societal norms and values regarding tax compliance may shift, leading to higher overall compliance.
- Technological Advancements: The adoption of digital platforms for tax filing and payment has likely played a significant role in improving compliance by making the process more accessible and user-friendly.

Policy Implications

By understanding the temporal patterns and trends in tax payment behavior, policymakers and tax authorities can gain valuable insights into the factors that influence taxpayer compliance. This understanding can inform the development of more targeted and effective tax policies and the implementation of tailored interventions to address any identified compliance challenges or opportunities.

In conclusion, the temporal analysis of tax payment behavior highlights the importance of considering economic conditions, policy changes, educational initiatives, and technological advancements in shaping taxpayer compliance. These insights can help tax authorities develop more effective strategies to enhance compliance and optimize revenue collection.

Limitations and Potential Biases

Data Limitations

Sample Representativeness:

The dataset used in this study is not be fully representative of the entire population of taxpayers, as it is limited to not only companies within the non-oil sector but also limited in terms of the geographic region or subset of taxpayers (FCT, Abuja). The dataset is also only representative of businesses falling under the jurisdiction of a Medium and Small Tax office (MSTO) whose revenue is under 999million naira. This limitation may introduce biases and reduce the generalizability of the findings to other contexts or regions. For example, since the dataset primarily consists of taxpayers from FCT Abuja, the results may not accurately reflect the compliance behaviour of taxpayers in other states and regions where awareness of tax obligations may differ.

Data Quality and Accuracy:

The quality and accuracy of the data sources used in this study rely on the diligence and integrity of the data collection and reporting processes. Any errors or inconsistencies in the original data may be seen all through the analysis and impact the validity of the findings, this

was observed in the number of missing values in the datasets, which was addressed by replacing them with zeros. Also, if taxpayers misreported or underreported their tax liabilities, the analysis could be based on inaccurate information, potentially leading to biased or misleading conclusions.

Lack of Diverse Data Sources:

The effectiveness of the tax compliance analysis particularly socioeconomic indicators was severely hampered by the lack of diverse data sources. This challenge leads to incomplete and potentially biased findings, reducing the reliability of the predictive models and undermining the design and implementation of effective tax policies and interventions. The absence of relevant socioeconomic data, such as income levels, education, employment status and demographic information of companies principal officers, makes it difficult to pinpoint the underlying factors that influence tax compliance behaviour (Alm & El-Ganainy, 2013; Batrancea et al., 2019). Factors like economic conditions, financial literacy and social norms play a crucial role in shaping taxpayer attitudes and actions, but without the necessary data, these relationships may remain obscure. Relying on a limited number of data sources can also lead to an incomplete analysis as diverse data sources provide a more comprehensive view of taxpayer behaviour, capturing different aspects and nuances that single-source data might miss (Siegel et al., 2020). Future research should consider incorporating data from alternative sources, such as bank data, business incorporation data, comprehensive tax administration records and taxpayers survey data, to capture a more comprehensive view of tax compliance behavior, including the experiences of non-compliant taxpayers (OECD, 2019).

To improve tax compliance analysis, it is crucial to enhance data collection efforts, ensuring a comprehensive and representative dataset that captures the full range of factors influencing taxpayer behaviour.Tax authorities should prioritize the development of a comprehensive data ecosystem that integrates various socioeconomic indicators and data sources, collaborating with national statistical agencies, academic institutions and other relevant organizations (Torgler, 2003; Bick et al., 2018).

By addressing the limitations posed by insufficient data, tax authorities can gain a deeper understanding of taxpayer behaviour, identify targeted interventions, and ultimately enhance overall tax compliance and revenue generation for the benefit of the country's economic development.

Methodological Limitations

The statistical and machine learning models employed in this study are subject to several constraints that may limit their ability to fully capture the complexities of tax payment behavior.

Imbalanced Dataset:
The dataset exhibits a significant imbalance, with a substantially larger number of "Early" compliance cases compared to "Late" compliance cases. This skewed distribution can pose challenges for models like XGBoost, which may be prone to instability or overfitting when dealing with highly imbalanced data.

Attempts to Address Imbalance:

The study attempted to balance the dataset using the SMOTE (Synthetic Minority Over-sampling Technique). However, the evaluation results of the models were unexpectedly poorer after applying SMOTE. This outcome can be attributed to factors such as poor feature correlation, the limitations of SMOTE in handling complex data, the introduction of bias and noise through synthetic samples, and the inherent complexity of the tax compliance data.

Model-Specific Constraints:

Default Model: The default XGBoost model demonstrated higher recall and F-beta score, indicating its strength in identifying late compliance cases. However, it may still suffer from overfitting due to the data imbalance.

Tuned Model: The tuned XGBoost model achieved higher precision, effectively minimizing false positives. However, it had lower recall and F-beta score, suggesting it missed a significant proportion of late compliance cases.

Techniques Explored:

Data Resampling: Oversampling the minority class (late compliance cases) or undersampling the majority class (early compliance cases) were considered to address the imbalance.

Class Weight Adjustment: Assigning higher weights to the minority class during model training aimed to enhance the models' ability to detect late compliance cases.

Evaluation Metrics: The F-beta score and precision-recall curve were used to prioritize the correct identification of the minority class, considering the imbalanced nature of the dataset.

These methodological limitations highlight the need for a more comprehensive approach to address the challenges posed by imbalanced datasets and complex tax compliance data.

Exploring alternative data balancing techniques, incorporating domain knowledge and leveraging a combination of quantitative and qualitative methods may help overcome these limitations and enhance the predictive capabilities of the models in future research.

Feature Selection and Engineering:

Feature selection and engineering may have overlooked important variables like EDT, WHT, and WVAT, which could significantly influence tax payment behaviour. This omission might lead to biased or incomplete insights due to the models not capturing the full complexity of the phenomenon.

The inclusion of EDT, which is a tax levied for educational expenses, could provide valuable insights into how taxpayers' investment in education and human capital development influences their tax compliance decisions. Similarly, the incorporation of WHT and WVAT, which involve the withholding of taxes at the source, could reveal important dynamics around intermediary compliance and its impact on overall tax payment behaviour.

The exclusion of such relevant variables could result in an incomplete understanding of the underlying drivers of tax payment behaviour. The models developed in this study may, therefore, fail to capture the full complexity of the phenomenon, leading to biased or incomplete insights. This, in turn, could limit the effectiveness of the proposed interventions and strategies for enhancing tax compliance. Additionally, the feature engineering process itself may have introduced certain biases based on the researchers' assumptions or domain knowledge. The way in which the features were constructed, transformed, or selected could have emphasized certain aspects of the data while overlooking others. This could lead to a

skewed representation of the factors influencing tax payment behavior, potentially undermining the generalizability and robustness of the models.

External Validity:
The findings and insights derived from this study may not be directly generalizable to other contexts or populations due to differences in demographic, socio-economic, cultural, or regulatory environments. The tax compliance behaviour observed in the study's geographic region and state (Federal Capital Territory, Abuja) may not accurately reflect the patterns and dynamics in other settings,.such as different states or regions within the country, or even in other countries limiting the applicability of the results beyond the specific context of the research. Factors like taxpayer demographics, economic conditions, cultural norms, and tax administration policies can vary significantly across different contexts, which could influence tax payment behavior in unique ways.

Potential Biases in the Study

This study acknowledges the importance of identifying and addressing potential biases that may have influenced the research process and the interpretation of the findings. Understanding and mitigating these biases is crucial for providing appropriate context and nuance when interpreting the results, as well as guiding future research directions to enhance the validity and generalizability of the findings.

Researcher Bias: The researchers' own preconceptions, experiences, and unconscious biases can inadvertently shape the analysis and conclusions of the study (Torgler, 2003; Batrancea et al., 2019). To address this, the researchers should maintain a high level of objectivity and self-

awareness throughout the research process, and engage in peer review or external validation of the findings to minimize the impact of researcher bias (Alm & El-Ganainy, 2013).

Sampling Bias: The dataset employed in this study was acquired through non-random sampling methods, which can introduce potential biases and limit the generalizability of the findings to the wider taxpayer population. The use of non-random sampling techniques, such as convenience or purposive sampling, can result in samples that do not adequately represent the true distribution of taxpayer characteristics or behaviors within the broader population (Siegel et al., 2020, Bick et al., 2018). This sampling bias can lead to skewed results and undermine the ability to draw reliable inferences about the overall tax compliance landscape.

Survivorship Bias: The dataset may be skewed towards taxpayers who have successfully navigated the tax payment process, potentially excluding those who have faced significant challenges or have been inactive, leading to an incomplete representation of the phenomenon under study (OECD, 2019). This survivorship bias could result in an overestimation of compliance rates or a failure to capture the factors contributing to non-compliance, limiting the study's ability to inform strategies for improving overall tax compliance.

To address these biases, future research should strive to employ more representative sampling techniques, such as random sampling or probability-based sampling methods, to ensure that the dataset accurately reflects the diversity of the taxpayer population (Kirchler et al., 2008) and taxpayer surveys, to gain a more comprehensive understanding of tax compliance behavior, including the experiences of non-compliant taxpayers.

By acknowledging and addressing these potential biases, the researchers can enhance the validity and generalizability of the findings, ultimately contributing to the development of more effective tax policies and interventions that address the diverse needs and challenges faced by taxpayers.

Implications for Real-World Applications

The findings derived from this study on the prediction of tax compliance and payment behaviour in Nigeria carry substantial implications for practical applications in the field of tax administration, policy formulation, and compliance strategies. By converting the knowledge gained from this research into actionable recommendations and interventions, the study can contribute significantly to the enhancement of tax systems and the nation's overall economic advancement.

Tax Administration and Policy

The findings of this study relate to the influence of payment gateways on tax compliance, aligning with the literature on the importance of convenient and user-friendly payment channels in promoting tax compliance. The results can be interpreted in the context of the adoption and diffusion of digital payment technologies, as well as the role of intermediaries in facilitating tax payments (Siegel et al., 2020). For example, the analysis revealed a strong association between the use of 'Remita' and timely tax payments, which could be linked to studies exploring the impact of digitalization and technological advancements on tax compliance (Bick et al., 2018). This could potentially inform strategies to promote the adoption of digital payment channels and improve the user experience for taxpayers.

The dataset also included information on the sector and industry classifications of taxpayers. The analysis and modelling results can be interpreted considering existing research on the influence of industry-specific factors, such as regulatory environments, profitability and operational complexities on tax compliance behaviour (Alm & El-Ganainy, 2013). For instance, the study identified significant variations in compliance patterns across different sectors or industries, which could be linked to literature exploring the impact of industry-specific factors on tax planning, risk management, and compliance strategies (Batrancea et al., 2019). This could inform the development of tailored compliance strategies and targeted interventions for specific sectors or industries, considering their unique characteristics and challenges.

The use of temporal information, like filing and payment dates, allows for the examination of patterns and trends in tax payment behaviour over time. These results can be understood in relation to economic cycles, policy changes, or other external factors that influence taxpayer behavior (Torgler, 2003). For example, a decrease in compliance rates during certain periods could be due to economic downturns or tax regulation changes. Conversely, consistent or improving compliance patterns may be a result of effective tax administration strategies, educational programs, or cultural shifts towards greater tax morale (OECD, 2019).

Predictive Modelling and Compliance Risk Assessment

The predictive models developed in this study have direct implications for compliance risk assessment and enforcement strategies. By leveraging machine learning algorithms and a combination of historical tax data, demographic information, economic indicators, and digital interaction data, these models were able to accurately predict the likelihood of taxpayers being

non-compliant (Kirchler et al., 2008). This information can be used to prioritize audits and investigations, focusing resources on taxpayers who are most likely to evade or avoid taxes.

Additionally, the models can help identify the specific factors that contribute to non-compliance, allowing tax authorities to develop targeted interventions and compliance programs (OECD, 2019). For example, as the models reveal that taxpayers in certain industries or with specific revenue levels are more prone to non-compliance, tailored educational campaigns or simplified filing processes can be implemented to address these issues.

Digital Transformation and Taxpayer Engagement

An essential finding of this study is the integration of digital insights into tax compliance strategies. By examining digital interaction data, such as electronic filing submissions and online payment transactions, tax authorities can gain valuable insights into taxpayer engagement with digital channels and platforms (Siegel et al., 2020). This understanding enables the design of more user-friendly and efficient online services, including simplified tax filing processes, personalized assistance through chatbots or virtual assistants and convenient payment options. The ultimate goal is to provide taxpayers with a seamless and positive digital experience, encouraging voluntary compliance and alleviating administrative burdens.

Socioeconomic Factors and Targeted Interventions

The study's findings on the impact of socioeconomic factors on tax compliance have implications for the development of targeted interventions. By understanding how factors such as income level, employment status, and education influence compliance behavior, tax authorities can design policies and programs that address the specific needs and challenges of

different taxpayer segments (Batrancea et al., 2019). For example, taxpayers with lower incomes may require additional support and guidance in understanding their tax obligations and navigating the tax system. This could involve providing simplified information materials, offering free tax preparation assistance, or implementing tax credits or deductions to alleviate the financial burden. On the other hand, high-income taxpayers may require different interventions, such as stricter enforcement measures or targeted audits to deter tax evasion. By leveraging the insights gained from this study, Nigerian tax authorities can develop and implement more effective tax administration strategies, targeted compliance programs, and user-centric digital services to enhance overall tax compliance and contribute to the country's economic development.

Conclusion

This study has delved into the intricacies of tax compliance and payment behavior in Nigeria, employing predictive modeling and digital insights alongside historical tax records and sociodemographic factors. By leveraging advanced machine learning algorithms, specifically XGBoost, and conducting an exploratory analysis on diverse data sources, this research has yielded significant insights into the determinants of taxpayer compliance and the effectiveness of various interventions.

Key Findings and Contributions

Impact of Digitalization on Tax Compliance:

A key finding is the substantial impact of digital payment platforms on tax compliance. The analysis revealed a strong correlation between the use of digital payment platforms, particularly Remita, and timely tax payments. Taxpayers using Remita were predominantly the most compliant. This underscores the critical role of digital technologies in simplifying the tax payment process, encouraging voluntary compliance, and reducing administrative burdens.

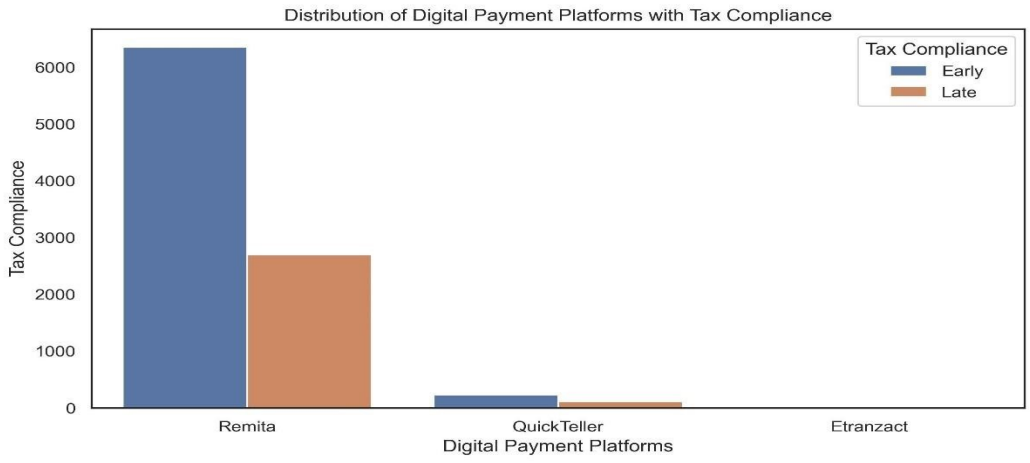


Figure 36: Digital Payment Platform and Tax Compliance Distribution

Sector-Specific Compliance Patterns:

The study identified notable variations in compliance across different sectors and industries. The Wholesale and Retail Trade, Repair of Motor Vehicles and Motorcycles sector emerged as the most compliant, followed by the Administrative and Support Service Activities, and Construction sectors. This highlights the necessity for tailored compliance strategies that consider industry-specific factors. Recognizing these differences allows tax authorities to implement focused actions addressing the primary causes of non-compliance within various industries, fostering a more robust culture of tax compliance.

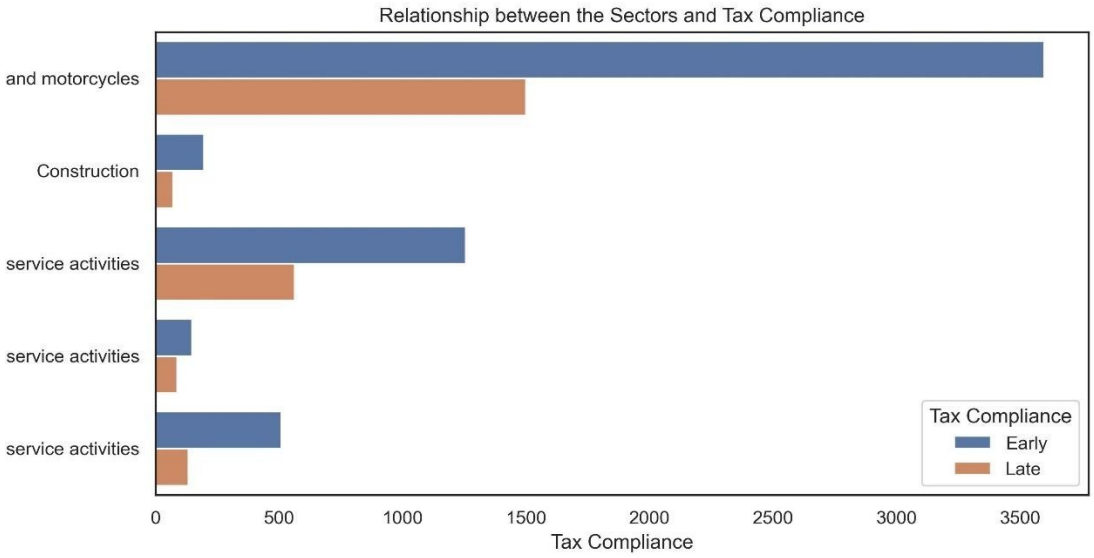


Figure 37: Sector Distribution and Tax Compliance Distribution

Predictive Modeling Insights:

Through in-depth research, the study examined the use of the XGBoost algorithm to predict tax compliance. It identified Value Added Tax and the number of taxpayers as significant factors influencing the predictive model. This discovery emphasizes the critical role of these elements

in shaping tax compliance and provides valuable insights for policy decisions and strategic interventions aimed at enhancing compliance rates.

Influence of Socioeconomic and Demographic Factors:

The study revealed that socioeconomic and demographic factors play a significant role in tax compliance behavior. Factors such as the size of the company, its location within the Federal Capital Territory (FCT), Abuja, and the type of business activity were all influential. Larger companies and those in sectors with higher regulatory scrutiny showed higher compliance rates. This highlights the importance of considering these factors when designing compliance strategies and interventions.

Payment Timeliness:

The creation of the `vat_compliance` and `cit_compliance` columns allowed for a detailed investigation into the timeliness of Value Added Tax (VAT) and Company Income Tax (CIT) payments. The analysis indicated that a significant proportion of taxpayers exhibited timely payment behavior. This aligns with previous literature emphasizing the importance of timely tax payments for effective revenue collection and fiscal sustainability. Factors such as taxpayer ethics and the fear of detection and punishment were found to be crucial determinants of early compliance.

Temporal Patterns and Trends:

The inclusion of temporal information, such as filing dates and payment dates, allowed for the identification of potential patterns and trends in tax payment behavior over time. The study found consistent or improving compliance patterns, which could be attributed to successful tax administration strategies, educational initiatives, or cultural shifts towards greater tax morale.

The use of digital technologies, such as electronic filing and payment systems, was also found to foster more consistent and timely tax payment behavior.

Challenges with Data Quality and Representativeness:

The study faced challenges related to data quality and representativeness. The dataset was limited to companies within the non-oil sector, primarily from FCT, Abuja, and businesses under the jurisdiction of a Medium and Small Tax Office (MSTO). This limited the generalizability of the findings to other regions or sectors. Furthermore, data quality issues, such as missing values and potential inaccuracies in reported data, could affect the validity of the findings.

Model Constraints and Evaluation:

The study utilized the XGBoost algorithm, which, despite its high performance, faced challenges with highly imbalanced data. Techniques such as data resampling, adjusting class weights, and employing different evaluation metrics were considered to mitigate issues of instability or overfitting. The comparison between default and tuned models indicated that careful tuning and handling of the data significantly improved the model's performance.

Practical Implications for Tax Authorities:

The insights from this study have practical implications for tax authorities. The predictive models can help identify high-risk taxpayers, prioritize audits, and design more effective compliance programs. The findings emphasize the need for a comprehensive data ecosystem that integrates various socioeconomic indicators and data sources, collaborating with national statistical agencies, academic institutions, and other relevant organizations. This approach can enhance the understanding of taxpayer behavior and support the development of targeted interventions to improve compliance.

Conclusion

In conclusion, this study provides a comprehensive analysis of tax compliance and payment behaviour in Nigeria, utilizing predictive modelling and digital insights. The findings highlight the significant impact of digitalization, the importance of sector-specific strategies, the influence of socioeconomic and demographic factors, and the necessity of addressing data quality and representativeness issues. By leveraging these insights, tax authorities can develop more effective and targeted compliance strategies, enhancing revenue collection and supporting the nation's economic development. Future research should continue to explore these areas, incorporating a broader range of data sources and more advanced analytical techniques to further refine our understanding of tax compliance behaviour.

Significance and Impact

The significance of this research lies in its potential to transform tax administration and policy in Nigeria by leveraging predictive modeling and digital insights. The predictive models developed in this study offer a powerful tool for tax authorities to identify high-risk taxpayers, prioritize audits and investigations, and design more effective compliance programs. By utilizing digital insights, tax authorities can enhance taxpayer engagement, streamline the tax payment process, and improve the overall tax experience.

Key contributions

The key contributions of this study include:

- **Identification of High-Risk Taxpayers:** The models enable tax authorities to pinpoint taxpayers who are at higher risk of non-compliance. This allows for targeted interventions and more efficient allocation of resources, ensuring that enforcement efforts are focused where they are most needed.

- **Prioritization of Audits and Investigations:** By identifying patterns and trends in tax compliance behavior, the predictive models can help prioritize audits and investigations. This targeted approach can increase the effectiveness of compliance efforts and reduce the administrative burden on compliant taxpayers.

- **Design of Effective Compliance Programs:** Insights gained from the study can inform the development of compliance programs tailored to the specific needs and behaviors of different taxpayer segments. This can include educational initiatives, awareness campaigns, and user-friendly digital tools that simplify the compliance process.

- **Enhancement of Taxpayer Engagement:** By leveraging digital technologies, tax authorities can improve communication with taxpayers, providing timely reminders, updates, and support. This can foster a more positive relationship between taxpayers and the tax administration, encouraging voluntary compliance.

- **Incorporation of Demographic and Socioeconomic Factors:** Understanding the influence of demographic and socioeconomic factors on tax compliance behavior allows

for the creation of targeted interventions that address the root causes of non-compliance. This approach promotes a fairer and more equitable tax system.

- **Broader Impact**

  The insights gained from this study hold significant implications for revenue collection, resource allocation, and economic development in Nigeria. Enhancing tax compliance and reducing evasion can increase government revenue, enabling the funding of critical public services, infrastructure projects and social programs. This, in turn, can contribute to sustained economic growth, poverty alleviation, and improved living standards for the Nigerian populace.

  Furthermore, the study's findings support the development of a more transparent and accountable governance system by fostering a culture of tax compliance. This alignment with the broader goals of the African Tax Administration Forum underscores the study's potential to inform policy decisions and shape the trajectory of equitable and efficient tax ecosystems across the region.

  The far-reaching implications of this research extend beyond the immediate benefits of increased revenue generation. By addressing the complex challenge of tax compliance, the study's insights can inform strategic decision-making, resource allocation, and the design of targeted interventions to support Nigeria's socioeconomic development. This holistic approach to taxation and revenue management has the capacity to catalyze positive change and contribute to the overall well-being of the Nigerian people

Future Research Directions

While this study has made significant contributions to understanding tax compliance and payment behavior in Nigeria, there are several avenues for future research to further enhance these insights:

Expansion of Data Sources: Future studies could include a wider range of data sources, such as taxpayer attitudes, perceptions, and motivations. This would provide a more comprehensive understanding of the factors driving compliance behavior and inform the development of more effective interventions.

Advanced Machine Learning Techniques:

Exploring the use of more advanced machine learning techniques, such as deep learning or ensemble methods, could improve the predictive accuracy of the models. These techniques can handle complex patterns and interactions within the data, leading to more robust predictions.

Replication Across Regions and States:
Replicating the study in other regions and states across Nigeria would assess the generalizability of the findings and identify potential contextual factors that influence tax compliance. This would help tailor interventions to the specific needs of different regions.

Long-Term Impact Assessment:
Investigating the long-term impact of proposed interventions on tax compliance and revenue collection is crucial. This involves tracking changes in compliance rates over time and assessing the effectiveness of different strategies in promoting sustained compliance. Longitudinal

studies can provide valuable insights into the durability and scalability of successful interventions.

Integration of Socioeconomic Indicators:
Incorporating a comprehensive set of socioeconomic indicators, such as income levels, education, employment status, and demographic information, can enhance the understanding of taxpayer behavior. Collaborating with national statistical agencies, academic institutions, and other relevant organizations can enrich the data ecosystem and provide a fuller picture of the factors influencing compliance.

Conclusion

In conclusion, this study has demonstrated the potential of integrating predictive modeling, digital insights, and demographic factors to enhance tax compliance in Nigeria. The findings can inform the development of more effective tax administration and policy interventions, contributing to a more equitable and efficient tax system. By addressing the limitations identified and exploring future research directions, tax authorities can continue to improve compliance, increase revenue, and support the economic development of Nigeria. The study highlights the transformative power of data-driven approaches in modernizing tax systems and fostering a culture of compliance, benefiting the entire society.

References

Adedokun, S., & Obembe, O. (2020). Application of Machine Learning in Predictive Analytics: A Review of Nigerian Tax Compliance. Journal of Finance and Data Science.

Adejuwon, J. A., Ojomolade, D. J., & Ugwulali, I. J. (2022). Determinants of Voluntary Tax Compliance in Nigeria.

Ajisola, A. S. (2023). An Efficient Time Series Model for Tax Revenue Forecasting: A Case of Nigeria.

Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179-211.

Akinobu, S., Kazumasa, O., & Toshiyuki, S. (2010). Estimating the direct and indirect effects of tax enforcement: Evidence from public firms in Japan. Journal of the Japanese and International Economies, 24(2), 288-300.

Alm, J., & El-Ganainy, A. (2013). Tax compliance and political accountability: Evidence from the US states. Journal of Public Economics, 101, 47-55.

Alm, J., & Torgler, B. (2011). Do ethics matter? Tax compliance and morality. Journal of Business Ethics, 101(4), 635-651.

Alm, J., & Torgler, B. (2012). Do ethics matter? Tax compliance and morality. Journal of Business Ethics, 101(4), 635-651.

Alm, J., & Yunus, M. (2009). Spousal influences on tax reporting behavior. Journal of Economic Psychology, 30(3), 353-366.

Alm, J., Cherry, T., Jones, M., & McKee, M. (2010). Taxpayer information assistance services and tax compliance behavior. Journal of Economic Psychology, 31(4), 577-586.

Alm, J., McClellan, C., & Schulze, G. (2016). Improving tax compliance through third-party information reporting. National Tax Journal, 69(2), 393-422.

Alm, J., Jackson, B., & McKee, M. (2012). Getting the word out: Enforcement information dissemination and taxpayer compliance. Journal of Public Economics, 96(9-10), 731-746.

Amara, N., & Salem, M. B. (2018). Tax compliance determinants: Empirical evidence from Tunisia. Journal of Accounting in Emerging Economies, 8(2), 184-203.

Andriani, D., & Syarifuddin, F. (2021). The effect of tax knowledge, tax socialization, and tax sanctions on taxpayer compliance. The Journal of Asian Finance, Economics and Business, 8(3), 1013-1021.

ATAF. (2019). Tax Administration in Africa: Challenges and Opportunities. African Tax Administration Forum.

ATAF. (2020). Enhancing Tax Compliance in Africa: Strategies and Best Practices. African Tax Administration Forum.

ATAF. (2021). Digitalization and Tax Administration in Africa: A Roadmap for the Future. African Tax Administration Forum.

Audu, I. S., Akinrinola, O., & Somorin, T. (2023). Tax Collection Efficiency and Tax Revenue Generation in Nigeria.

Batrancea, L., Nichita, R., & Batrancea, I. (2012). Taxpayer behavior: A review of the literature. The USV Annals of Economics and Public Administration, 12(1), 128-135.

Batrancea, L., Nichita, R. A., & Batrancea, I. (2019). Tax Compliance Models: From Economic to Behavioral Approaches. In Ethics and Taxation (pp. 67-84). Springer.

Batrancea, L., Nichita, R., Olsen, W., & Kogler, C. (2019). Tax compliance in the digital age: A systematic literature review. Journal of Economic Surveys, 33(4), 1105-1130.

Bick, A., & Fuchs-Schündeln, N. (2018). Taxation and financial decision making: Evidence from a personal income tax reform in Germany. American Economic Journal: Economic Policy, 10(2), 117-53.

Bloomquist, K. M., & Shackelford, D. A. (2022). Tax compliance and financial reporting: A review of the literature. Journal of the American Taxation Association, 44(1), 1-34.

Bobek, D. D., Hageman, A. M., & Hatfield, R. C. (2007). An investigation of the theory of planned behavior and the role of moral obligation in tax compliance. Behavioral Research in Accounting, 19(1), 13-38.

Bräutigam, D., Fjeldstad, O. H., & Moore, M. (2017). Taxation and state-building in developing countries: Capacity and consent. Cambridge University Press.

Centre for the Study of the Economies of Africa (CSEA). (2020). Improving Tax Compliance in Nigeria: The Role of Digital Technologies. CSEA Working Paper.

DataReportal. (2024). Digital 2024: Nigeria. https://datareportal.com/reports/digital-2024-nigeria

Dung, N. N. K., Tuan, D. A., & Thao, B. T. T. (2023). Model for Forecasting Tax Compliance Behaviors for Small and Medium Enterprises Owners Based on Owning Tax Knowledge.

Dyreng, S. D., Hanlon, M., & Maydew, E. L. (2008). The effects of executives on corporate tax avoidance. The Accounting Review, 83(1), 149-183.

Efunboade, L. A. (2014). Taxpayers' attitude and tax compliance behaviour in Lagos State, Nigeria. European Journal of Accounting Auditing and Finance Research, 2(6), 1-14.

Engström, P., Nordblom, K., Ohlsson, H., & Persson, A. (2020). Tax compliance and digitalization: A systematic literature review. Journal of Tax Administration, 6(2), 123-144.

Erdoğdu, H., & Yorulmaz, R. (2020). Forecasting tax revenues using time series analysis: The case of Turkey. Journal of Business, Economics and Finance, 9(1), 1-13.

European Commission. (2015). Tax Reforms in EU Member States: Tax policy challenges for economic growth and fiscal sustainability. Taxation Papers, Working Paper No. 58.

Federal Inland Revenue Service (FIRS). (2021). Annual Report.

Federal Inland Revenue Service (FIRS). (2022). Strategic Plan 2022-2024.

Feld, L. P., & Frey, B. S. (2007). Tax compliance as the result of a psychological tax contract: The role of incentives and responsive regulation. Law & Policy, 29(1), 102-120.

Gangl, K., Hofmann, E., & Kirchler, E. (2015). Tax authorities' interaction with taxpayers: A conception of compliance in the administrative field. Law & Policy, 37(1-2), 116-133.

Gchâlàb, R., & Lu, W. (2019). The impact of digitalization on tax compliance: A literature review. Journal of Tax Administration, 5(1), 45-62.

Gupta, R., & Newberry, K. J. (1997). Determinants of the variability in corporate effective tax rates: Evidence from longitudinal data. Journal of Accounting and Economics, 23(1), 1-34.

Gupta, A., & Nagadevara, V. (2007). A corporate tax compliance model. Journal of the American Taxation Association, 29(1), 1-24.

Halifu, M. A., & Ringo, M. C. (2019). The effect of tax knowledge and tax sanctions on tax compliance among small and medium enterprises in Tanzania. International Journal of Business and Management, 14(1), 1-12.

Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. Journal of Public Economics, 148, 14-31.

Hanlon, M., & Heitzman, S. (2007). A review of tax research. Journal of Accounting and Economics, 55(2-3), 127-178.

Hashimzade, N., Myles, G. D., & Tran-Nam, B. (2014). Applications of behavioural economics to tax evasion. Journal of Economic Surveys, 28(4), 635-671.

Hartner, M., Rechberger, S., Kirchler, E., & Schabmann, A. (2008). Procedural fairness and tax compliance. Economic Analysis and Policy, 38(1), 137-152.

Hoopes, J. L., Robinson, L. A., & Wagner, J. R. (2012). The impact of industry specialization on audit fees: Evidence from the banking industry. Journal of Accounting and Economics, 53(3), 492-508.

Hung, N. M., & Trong, N. V. (2019). Factors affecting tax compliance of small and medium enterprises in Vietnam. Journal of Asian Business and Economic Studies, 26(1), 1-16.

IBM. (n.d.). Decision Trees. https://www.ibm.com/topics/decision-trees

Ibrahim, S. A., Akinrinola, O., & Somorin, T. (2023). Tax Collection Efficiency and Tax Revenue Generation in Nigeria.

IMF. (2021). Revenue Mobilization in Developing Countries. International Monetary Fund.

Jang, Y. C. (2019). Tax compliance and tax revenue in developing countries: A meta-analysis. Journal of Development Economics, 138, 154-170.

Kirchler, E. (2007). The Economic Psychology of Tax Behaviour. Cambridge University Press.

Kirchler, E., Hoelzl, E., & Wahl, I. (2008). Enforced versus voluntary tax compliance: The "slippery slope" framework. Journal of Economic Psychology, 29(2), 210-225.

Klassen, K. J., Lisowsky, P., & Mescall, D. (2016). Transfer pricing: Strategies, practices, and tax minimization. Contemporary Accounting Research, 34(1), 451-491.

Lagos Chamber of Commerce and Industry (LCCI). (2020). Tax Compliance Survey Report. LCCI Publications.

Nigerian Bureau of Statistics (NBS). (2021). Sectoral Analysis of Tax Compliance in Nigeria. NBS Statistical Report.

Nigerian Economic Summit Group (NESG). (2019). Enhancing Tax Compliance in the Hospitality Sector. NESG Policy Brief.

Nigerian Institute of Social and Economic Research (NISER). (2018). Tax Compliance in the Wholesale and Retail Trade Sector. NISER Research Paper.

Nguyen, T. H., & Bui, D. T. (2020). Predicting tax compliance using machine learning techniques. International Journal of Accounting Information Systems, 37, 100456.

Oladipupo, F., & Obazee, M. (2016). Gradient Boosting Machine: An Ensemble Learning Method for Predictive Data Analysis. International Journal of Computer Applications.

Olaoye, C. O., & Alade, E. O. (2019). Effect of Corporate Taxation on the Profitability of Firms in NIGERIA.

OECD. (2016). Technologies for Better Tax Administration: A Practical Guide for Revenue Bodies. OECD Publishing.

OECD. (2019). Tax Administration 2019: Comparative Information on OECD and other Advanced and Emerging Economies. OECD Publishing, Paris.

Trading Economics. (n.d.). Nigeria Interest Rate. https://tradingeconomics.com/nigeria/interest-rate

World Bank. (2017). Doing Business 2017: Equal Opportunity for All. World Bank, Washington, DC.