BINDURA UNIVERSITY OF SCIENCE EDUCATION

**BINDURA UNIVERSITY OF SCIENCE EDUCATION**

**FACULTY OF SCIENCE AND ENGINEERING**

**DEPARTMENT OF STATISTICS AND MATHEMATICS**

# ANALYSIS OF MUNICIPAL REVENUE ARREAR PERFORMANCE – A DATA SCIENCE APPROACH: A CASE OF BINDURA MUNICIPALITY.

**BY**

**MASAIRE EMMANUEL**

**B224977B**

**A DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT OF THE BACHELOR OF SCIENCE HONOURS DEGREE IN STATISTICS AND FINANCIAL MATHEMATICS.**

**SUPERVISOR: MR B KUSOTERA**

**JUNE 2026**

## APPROVAL FORM

**This is to certify that this research project is the result of my own research work and has not been copied or extracted from past sources without acknowledgement. I hereby declare that no part of it has been presented for another degree in this University or elsewhere.**

**MASAIRE EMMANUEL**

**B224977B**

**Signature: _____**  **_____ Date: _____09/06/2026_**

**Certified by**

**MR B KUSOTERA**

**Supervisor**

**Signature: _____**  **_____ Date: _____09/06/2026**

**Dr M MAGODORA**

**Chairperson**

**Signature: __**  **_ Date: ____15/08/2026_____**

**Authorship Declaration Statement**

Title of the Thesis: ANALYSIS OF MUNICIPAL REVENUE ARREAR PERFORMANCE – A DATA SCIENCE APPROACH: A CASE OF BINDURA MUNICIPALITY

Author: MASAIRE EMMANUEL

Program: Honours Bachelor of Science Degree in Statistics and Financial Mathematics

I, the undersigned author of the above-mentioned thesis, hereby declare that:

1. This thesis is my original work and has been prepared by me in accordance with the institution's requirements.
2. All sources, data, and references used in this thesis have been acknowledged and cited appropriately.
3. This thesis has not been submitted elsewhere for any degree or diploma.
4. I have obtained all necessary permissions for the inclusion of third-party content where applicable.

I affirm that this declaration is made with full integrity and in compliance with the institution's policies and academic practice.

Author's Signature: 

Date: 09/06/2026

**ACKNOWLEDGEMENTS**

I would like to thank my parents Mr and Mrs Masaire for the unwavering support and love in my education journey. Their encouragement and support towards my studies have brought me this far. I would like to acknowledge my supervisor Mr B Kusotera for always pushing me not to give up. I would like to thank God the Almighty foe keeping me through this whole journey and affording me his grace to complete my education at BUSE.

## ABSTRACT

In this research, we examine the municipal revenue arrear performance for Bindura Municipality using data science, specifically on the prediction of customer arrears behavior. Municipalities in Zimbabwe are facing serious sustainability issues in terms of revenue generation due to high rates of defaults, poor billing practices and limited application of predictive analytics techniques. Although there is abundance of information related to billing and payments from customers, very little research has been done in applying machine learning in order to predict municipal revenue arrear performance. Based on 1,499 municipal customers, this research used three classification models: Logistic Regression, Random Forest, and XGBoost in order to predict whether the customer had arrears. Dependent Variable is Revenue Status (Arrear = 1, No Arrear = 0) while Independent Variables are Billing Amount, Property Type, Payment Type, and Employment Status. The process involved data pre-processing (imputation, scaling, outlier analysis and stratified sampling) then evaluation was based on Accuracy, Sensitivity, Specificity, Precision, F1-Score, and AUC. The findings show that although Logistic Regression model demonstrated statistically significant p-values (<0.05), its predictive capacity was poor since sensitivity was extremely low (3.9%) and the model did not display good discriminatory power with an AUC of 0.538. On the other hand, machine learning models proved superior in predictive accuracy where Random Forest attained an AUC of 0.901 through balancing sensitivity (81.78%) and specificity (79.48%). However, XGBoost proved to be the most effective model with 91.33% accuracy, 90.58% arrears recall and AUC of 0.958. From the findings, arrears are determined by complex non-linear relationship and hence the need for the adoption of ensemble learning. The study recommends the use of XGBoost model in managing municipal revenue arrear performance by employing risk based revenue strategy, target enforcement approach and financial planning by municipality.


### Recommendations

1. Apply XGBoost technique to predict municipal arrears.
2. Develop a risk-based revenue management framework using XGBoost as the underlying tool.
3. Design early intervention mechanism targeting medium- and high-risk accounts.
4. Modernize municipal financial infrastructure with regard to data collection and management.
5. Reconduct and validate the model regularly to retain predictive efficiency.

Contents

APPROVAL FORM ..... ii

Authorship Declaration Statement ..... iii

ACKNOWLEDGEMENTS ..... iv

ABSTRACT ..... vi

CHAPTER 1 ..... 1

1.0 INTRODUCTION ..... 1

1.1 BACKGROUND OF THE STUDY ..... 1

1.2 STATEMENT OF THE PROBLEM ..... 2

1.3 RESEARCH OBJECTIVES ..... 2

1.4 RESEARCH QUESTIONS ..... 2

1.5 SCOPE OF THE STUDY ..... 3

1.6 SIGNIFICANCE OF THE STUDY ..... 3

1.7 ASSUMPTIONS OF THE STUDY ..... 3

1.8 LIMITATIONS OF THE STUDY ..... 3

1.9 DEFINITION OF KEY TERMS ..... 3

1.10 CONCLUSION ..... 4

CHAPTER 2 – LITERATURE REVIEW ..... 5

2.0 INTRODUCTION ..... 5

2.1 THEORETICAL LITERATURE REVIEW ..... 5

2.1.1 Theories of Customer Default and Arrears ..... 5

2.1.2 Theories of Billing and Revenue Collection Efficiency ..... 6

2.1.3 PUBLIC FINANCE THEORY ..... 7

2.1.4 Theories of Taxation ..... 7

BINDURA UNIVERSITY OF SCIENCE EDUCATION

2.1.5 The Contributing Role of Data Science in Public Finance ..... 8

2.1.6 Governance and Institutional Theories ..... 8

2.1.7 Challenges Facing Local Authorities in Raising Revenues ..... 8

2.2. Conceptual Framework ..... 9

2.2.1 Components of the Conceptual Framework ..... 9

2.2.2 Conceptual Relationship Model .....11

2.2.3 Hypothesis Development .....11

2.2.4 Conceptual Contribution of the Study .....11

2.2.3 Visual Representation of the Conceptual Framework..... 12

2.2.4 Implications of the Conceptual Framework ..... 13

2.2.5 Conclusion of the Conceptual Framework ..... 13

2.3 Empirical Literature ..... 13

2.3.1 Techniques for Revenue Generation ..... 13

2.3.2 Data Science Applications in Governance ..... 16

2.3.4 Challenges in Data-Driven Revenue Management ..... 17

2.3.5 Summary of Empirical Findings ..... 18

2.4 Literature Synthesis & Research Gap ..... 18

2.4.1 Synthesis of Theoretical and Empirical Literature ..... 18

2.4.2 Research Gap Identity ..... 19

2.4.3 Implications for the current study ..... 20

2.5 Summary of Key Findings ..... 20

2.5.1 Connotations for Future Studies ..... 20

2.5.2 Getting Ready for the Next Chapter ..... 21

2.6 Conclusion ..... 21

BINDURA UNIVERSITY OF SCIENCE EDUCATION

CHAPTER 3: RESEARCH METHODOLOGY ..... 22
3.0 Introduction ..... 22
3.1 Research Design ..... 22
3.2 Data Collection Techniques ..... 23
3.4 Data cleaning and Pre-processing ..... 23
3.5 Variable description ..... 25
3.5.1 Description of variables - Table 3.1 ..... 25
3.5.2 Variable justification ..... 25
3.6 Models ..... 26
3.6.1 Logistic regression ..... 26
3.6.2 Random Forests (RF) ..... 29
3.6.3 XGBoost ..... 34
3.7 Model Evaluation and Selection ..... 34
3.7.1 Evaluating forecasting models ..... 34
3.7.2 Metrics of forecast/model performance ..... 34
3.7.3 Classification performance measures: ..... 35
3.7.4 Model comparison and testing ..... 35
3.7.5 Model explainability ..... 35
3.7.6 Practical considerations for deployment ..... 35
3.8 Chapter Summary ..... 35
CHAPTER 4: DATA PRESENTATION, ANALYSIS AND DISCUSSION ..... 37
4.1 Introduction ..... 37
4.2 Descriptive Statistics ..... 37
4.2.1 DESCRIPTIVE SUMMARY STATISTICS ..... 37

4.2.2 Descriptive Statistics Interpretation ..... 38

4.2.3 Implications for Modeling ..... 39

4.3 DIAGNOSTICS TESTS ..... 40

4.3.1 Outlier Detection ..... 40

4.3.2 Correlation Matrix Analysis ..... 41

4.3.3 Variance Inflation Factor (VIF) Test ..... 41

4.3.4 Sample Size Adequacy ..... 42

4.3.5 Summary of Pre-Estimation Tests..... 42

4.4 MODEL BUILDING ..... 43

4.4.1 Logistic Regression ..... 43

4.4.2 RANDOM FOREST MODEL ..... 46

4.4.3 XGBOOST MODEL ..... 48

4.5.1 Summary of Model Performance ..... 50

4.5.2 Revenue Forecasting ..... 51

CHAPTER 5: DISCUSSION, POLICY IMPLICATIONS AND CONCLUSION ..... 55

5.1 Introduction ..... 55

5.2 Discussion of Key Findings ..... 55

5.3 Policy Implications ..... 56

5.3.1 Targeted Revenue Enforcement ..... 56

5.3.2 Risk Based Revenue Management ..... 56

5.3.3 Evidence-Based Decision Making ..... 56

5.3.4 Institutional Modernization ..... 57

5.4 Theoretical Implications ..... 57

5.5 Study Limitations ..... 57

5.6 Recommendations ..... 57

BINDURA UNIVERSITY OF SCIENCE EDUCATION

5.7 Chapter Conclusion ..... 58

Reference ..... 59

# TOPIC: ANALYSIS OF MUNICIPAL REVENUE ARREAR PERFORMANCE – A DATA SCIENCE APPROACH: A CASE OF BINDURA MUNICIPALITY

## CHAPTER 1

### 1.0 INTRODUCTION

Municipalities play a crucial role in local governance, particularly in resource allocation, infrastructure, public service delivery and economic development. In Zimbabwe, the financial sustainability of municipalities like Bindura is increasingly challenged by economic fluctuations and inefficient revenue collection systems. This study aims to analyze the revenue streams of Bindura Municipality using a data science approach which can provide valuable insights into optimizing revenue generation and management practices. Gunter A. (2018). Local Government Finance in Zimbabwe: Challenges and Opportunities, World Bank. (2021). Municipal Finance in Zimbabwe.

### 1.1 BACKGROUND OF THE STUDY

Bindura Municipality located in the Mashonaland Central Province, has historically relied on various revenue sources which include property tax, services fees and government grants for support. The Zimbabwe constitution gives municipalities the authority to impose rates and taxes to residents of the area both rural and urban. However, the effectiveness of these revenue streams has been inconsistent due to factors such as economic downturns and administrative inefficiencies. The integration of data science techniques can enhance understanding of revenue patterns and inform better decision making. According to Tonhodzai et al (2015:76), the majority of Zimbabwe municipalities are to provide services to their citizens and they depend on financial assistance from the government. As much as local government depend on the supreme authorities financial resources, the supreme authorities are incapable of providing municipalities with resources, so aggravating the municipalities prediction.

According to Mudyanadzo and Nzwatu (2018:7) Zimbabwean municipalities still have a long way to go before being considered financially sustainable and capable of carrying out their duties as effective and efficient councils that deliver quality services at a reasonable cost. Since municipal delivery non-payment has been a frequent occurrence throughout Zimbabwe, the municipalities overall financial situation has been exacerbated by lack of revenue.

Payment default continued even after a multi-currency system was implemented in 2009 leaving the bleeding institutions unabated. The Minister of Local Government and Urban Development used the authorities given to them by the Urban Council Act (chapter 29:15) to second resuscitation teams in Bindura, Redcliff in Kwekwe and Chinhoyi, in response to a number of public outcry regarding subpar service delivery in various municipalities. Remarkably, neither the debtor`s ratio nor communities liquidity situation improved.

However, local government need to overcome obstacle related to revenue production in order to provide adequate service delivery objectives. Examining the effectiveness of revenue production in connection to service delivery commitments is crucial. Government grants were once given but later cancelled for local government that are able to support themselves.

## 1.2 STATEMENT OF THE PROBLEM

The municipality of Bindura is still having issues of revenue collection arising from the accruing of customer arrears. Delinquencies in municipal bills have a negative impact on service provision, maintenance of services and financial sustainability. Although there are customer billing records, there is limited use of predictive analytics to identify customers that are likely to default. Thus, the aim of this study is to create predictive models which can help the municipality to categorize the customers that are at risk and enhance the efficiency of revenue collection.

## 1.3 RESEARCH OBJECTIVES

- To identify factors affecting municipal revenue arrear performance.  To develop predictive models for municipal revenue forecasting.
- To provide data driven recommendations for revenue enhancement,

## 1.4 RESEARCH QUESTIONS

- What are the primary sources of municipal revenues and their related contributions over time?
- How do socioeconomic and environmental factors influence revenue variability and can these be modelled predictively?
- What data science techniques (e.g regression models, clustering or neural networks) are most effective for forecasting municipal revenue trends?
- How can insights from data analysis inform policy recommendations for improving revenue management in municipalities?

## 1.5 SCOPE OF THE STUDY

This study focuses on Bindura Municipality, analyzing its revenue arrear performance data over the past years (Bindura Municipality Financial Reports 2020 – 2024). It has various revenue sources, including property tax, service fees and other municipal charges. The study will employ data science approach such as statistical analysis, machine learning and data visualization.

## 1.6 SIGNIFICANCE OF THE STUDY

The research is significant for local government officials, policymakers, and stakeholders in Bindura Municipality. Employing data science techniques, the study aims to provide insights that can enhance revenue arrear performance and improve the municipality financial sustainability over the coming years. It contributes to the broader field of municipal finance and data driven governance.

## 1.7 ASSUMPTIONS OF THE STUDY

- The data provided by Bindura Municipality is accurate and reliable.
- Identified revenue sources will remain stable during the study period.
- Stakeholders will engage positively with the findings and recommendations.

## 1.8 LIMITATIONS OF THE STUDY

The limitation of the study is that it only focus on Bindura Municipality. According to Smith, R. (2019) Limitations in Public Sector Research, the extent of this inquiry is restricted by the quality and accessibility of the data gathered. The accuracy of the analysis could be impacted by inaccurate or insufficient data. Furthermore, as economic conditions and cultural norms may also be important, the study does not fully account for all factors that lead to municipal rate default.

## 1.9 DEFINITION OF KEY TERMS

**REVENUE** – refers to the money obtained through regular business operations, Hayes (2023)

**FINANCIAL SUSTAINABILITY** – (Govinaraju et al. (2015), is the ability of a project to raise the necessary funds to fulfil all of its obligations regarding resources and finances.

**DATA SCIENCE** – An interdisciplinary field that uses scientific methods, processes, and systems to extract knowledge and insights from data in various forms.

**REVENUE MANAGEMENT** – A collection of strategies and tactics that firms use to scientifically manage demand for their products and services (John M. McCoy (2001)).

**MUNICIPAL REVENUE** – is the income generated by local government entities from various sources including taxes and fees (Provost. F, Fawcett T. (2013).

### 1.10 CONCLUSION

This chapter has introduced the study on municipal revenues in Bindura Municipality, outlining the background, problem statement, objectives and significance of the research. The following chapters will delve deeper into the literature, methodology, findings and recommendations.

## CHAPTER 2 – LITERATURE REVIEW

### 2.0 INTRODUCTION

The following chapter discusses the current literature related to municipal revenues and application of data science methodologies. In particular, it gives a review of theoretical frameworks and empirical literature, and identifies the main gaps in the relevant knowledge field.

### 2.1 THEORETICAL LITERATURE REVIEW

Generating revenues and managing arrears in municipalities requires drawing from such academic disciplines as economics, behavioral theory, institutions, and administration. In order to understand the causes of default, creation of arrears, and inefficiencies in billing and collection process, it is

necessary to integrate several theoretical perspectives – public finance, taxation, compliance, and behavioral economics. This section highlights the theories relevant to the phenomenon of arrears and revenue collection in Bindura municipality.

### 2.1.1 Theories of Customer Default and Arrears

#### 1. Ability-to-Pay Theory

In case of customer defaults and arrears, several theories are used to understand why this problem occurs. One of such theories is the Ability-to-Pay Principle. According to the principle, the ability to make the necessary payments depends on the financial capabilities of the individual. Therefore, people with low incomes, unemployment, or cash problems have greater likelihood to fall into arrears. Empirical evidence from developing countries confirms the existence of positive correlations between poverty, unemployment, and unpaid municipal rates. Fluctuations in household incomes raise the likelihood of customer default. Hence, Bindura is not an exception, and the city is highly likely to experience economic instability, resulting in arrears among customers.

#### 2. Strategic Default Theory

The phenomenon of strategic default occurs when customers deliberately fail to make payment despite having the necessary funds, expecting the absence of severe consequences. Under this theory, the customers' reasoning behind the failure to make payments is rational, as weak penalties or inconsistent cutoffs mean that people see no point in making timely payments.

#### 3. Moral Hazard Theory

The moral hazard theory states that customers may fall into arrears if they believe that noncompliance will not result in any sanctions. Indeed, many jurisdictions refrain from taking stringent actions out of political considerations. In addition, the expectations of obtaining any kind of assistance from the government increase the probability of falling into arrears.

#### 4. Life-Cycle and Income Smoothing Theory

The Life-Cycle Hypothesis postulates that people try to smooth their consumption through life events. In case of an unexpected adverse event, the family will face the question of where the budget is spent. Essential items, including food and rent, take priority. This theory explains that arrears may emerge even in the case of customers who usually pay their bills, experiencing temporary hardships.

**5. Behavioral Economics and Payment Compliance**

Behavioral economics is concerned with psychological and social determinants of compliance. Factors like present-biased preferences, loss aversion, and procrastination have the ability to influence compliance. Under the Social Norms Theory, people are more willing to pay when the majority agrees that payments are expected. Hence, non-compliance becomes less common. It is an important aspect in considering the motives of municipalities' customers.

### 2.1.2 Theories of Billing and Revenue Collection Efficiency

Revenues generated by municipalities may depend on the efficiency of their billing systems.

**1. Administrative Efficiency Theory**

Under the admin efficiency theory, successful revenue generation is linked to the accuracy of billing, timing of the invoice issuance, clarity of statements, and correctness of documentation. Errors occurring in billing create distrust leading customers to postpone payments or neglect bills altogether. Besides, the use of clumsy manual processes typically results in higher expenses and a high error rate.

**2. Technology Adoption Theory**

Another important concept in discussing billing efficiency is the Tech Adoption theory. Based on Technology Acceptance Models, this theory emphasizes the role played by perceived usefulness and ease-of-use in customers' decision to switch to digital billing. Automation of processes, implementation of mobile payment solutions, and integration into financial management platforms significantly increase convenience, minimize errors, boost clarity, and improve compliance.
However, traditional billing systems are more likely to produce arrears.

### 2.1.3 PUBLIC FINANCE THEORY

Public finance is the foundation of all municipal revenue generating processes. It includes numerous principles associated with taxation, expenditures, and the role of government in general.

1. **Musgrave's Three Functions of Government:** Richard Musgrave claims that a government performs three functions which are, allocation, distribution, and stabilization. These fundamental principles can serve as a guidance in effective management of local municipality

resources and correction of any anomalies in equity. When analyzing revenue mechanisms in Bindura municipality, these functions should be kept in mind.

2. **Tiebout's Model of Local Public Goods:** Charles Tiebout suggests that people "vote with their feet" and prefer living in jurisdictions where preferences for certain public goods and taxation levels match best. Such model highlights the importance of local revenue generation as municipalities compete among themselves for customers and businesses.

### 2.1.4 Theories of Taxation

Taxation plays an important role in generating municipal revenue. Major theories associated with taxation include:

1. **Benefit Principle:** Benefit principle assumes the correlation between the benefits received and taxes paid. Therefore, the theory can be helpful in establishing a fair and appropriate taxation scheme.

2. **Ability-to-Pay Principle:** According to the ability-to-pay principle, tax rates should depend on the ability of the taxpayer. The theory calls for equality, ensuring that wealthier people pay more to municipality budget.

### 2.1.5 The Contributing Role of Data Science in Public Finance

Increasingly often, data science is applied to optimize revenue collection in municipalities.

1. **Decision-Making Models:** Data-driven decision making allows improving resource allocations and revenues by using predictive modeling techniques to make predictions about budgets and financial performance.

2. **Machine Learning in Revenue Optimization:** Machine learning techniques have been used to analyze historic data, offering more precise forecasts and improvements to tax compliance. Regression analysis and classification algorithms help detect behavioral patterns among taxpayers and provide recommendations for revenue collection.

### 2.1.6 Governance and Institutional Theories

1. **New Public Management**: This theory emphasizes efficiency, performance measurement, and accountability in the delivery of public service. NPM principles can guide municipalities in adopting data-driven strategies that enhance revenue management.

2. **Institutional Economics**: This framework focuses on the role of institutions in shaping economic behavior. It provides insight into how institutional arrangements could either facilitate or hinder effective revenue collection in municipalities.

### 2.1.7 Challenges Facing Local Authorities in Raising Revenues

The local governments face a number of challenges in enhancing revenues, which include:

- Resource limitation: The local governments are bound to have resource limitation, such as lack of human resources, that might limit the governments from undertaking all the duties involved in the collection of revenues.
- Arrears accounts: Most of the local governments have arrears accounts, especially taxes and fines that generate foregone revenues and additional costs incurred during the recovery process.
- Non-compliance: Certain individuals as well as business organizations are capable of intentionally avoiding the payment of charges/taxes thereby imposing a heavy toll on the process of collecting revenues on behalf of the local government.
- Regulatory flux: The regulations concerning taxation, as well as other regulatory systems, keep changing from time to time, and it is essential that the local governments are updated on these developments to ensure that the processes of raising revenues within these governments are precise and efficient.
- Resistance to collection: There might be people who would not wish to yield to collection, thus limiting the amount of revenues collected.

## 2.2. Conceptual Framework

This conceptual framework demonstrates how customer characteristics are linked to municipal arrears and revenue arrear performance. The performance of revenues in a city is measured by the presence or absence of arrears. This conceptual framework examines the influences of economic capacity, structural, and behavioral determinants on the revenue arrear performance. These include billing amount, property type, payment history, and employment status that influence a customer's tendency to be in arrears or not, respectively affecting the revenue arrear performance.

### 2.2.1 Components of the Conceptual Framework

An essential component of the conceptual framework includes independent variables, dependent variable, and moderating/contextual factors.

**1. INDEPENDENT VARIABLES**

**a. Billing Amount**

It denotes the fiscal obligation that is required from a customer. Based on Ability-to-Pay Theory, large billing amounts increase financial strain leading to arrears. Furthermore, under the Life-Cycle Theory, customers who are experiencing monetary problems are likely to opt-out of making the payments. Lastly, very large amounts of the billing raise fairness concerns, as per the Benefit Principle. All said and done, increased billing amounts increase the possibility of arrears, hence revenue arrear performance.

**b. Property Type**

The property type is the residential, commercial, or industrial structure used to distinguish differences in economic and structural characteristics of ratepayers. First, commercial properties are likely to provide customers with continuous and higher revenue income sources. Conversely, low-income residential property types face greater challenges in meeting their payment obligations. Building design and societal structure play a critical role in influencing payment capability. Therefore, we anticipate that revenue arrear performance will depend on the income continuity and billing structure.

**c. Payment History**

This measures how well a customer meets their payment obligations within a specified timeframe. Behavioral economics theory argues that payment behavior determines the likelihood of making payment decisions. Additionally, consistent failure to make payments may denote inadequacy in enforcement efforts as per Deterrence Theory. Moreover, customers are likely to build social norms regarding payment or lack thereof, as per the Social Norms Theory. Poor payment history increases future arrears, thus revenue arrear performance. To note, payment history emerges as the most critical determinant of arrears.

**d. Employment Status**

It defines economic capability of customers. As argued in Ability-to-Pay Theory, payment ability depends on income stability, and customers whose income fluctuates are likely to experience arrears. Additionally, unemployed and/or informally employed customers have irregular and unpredictable income streams, raising chances of failing to make timely payments. Unemployment and inconsistent employment status therefore increase chances of becoming customers in arrears, hence poor revenue arrear performance.

**2. Dependent Variable: Revenue arrear performance**

Here, revenue arrear performance is defined as the presence of absence of arrears. Zero represents the absence of arrears, thus revenue arrear performance, while one represents existence of arrears, thus poor revenue arrear performance. In a municipality, high arrears rates create financial instability, inaccurate budgets, and inability to offer services. Increased revenue arrear performance improves these dimensions as well.

**3. Moderating / Contextual Variables**

Moderating or contextual variables are likely to indirectly influence relationship between independent and dependent variables. Such variables may include enforcement policy, effectiveness of billing systems, effectiveness of services offered, and general governance effectiveness. These variables can potentially moderate either positively or negatively the effect of independent variables on the likelihood of arrears.

**2.2.2 Conceptual Relationship Model**

A conceptual model analyzes the determinants of revenue arrear performance including likelihood of arrears or payment backlog. This implies that the billing amount, property type, payment history, and employment status determine revenue arrear performance.

Simplified relationship;

Revenue arrear performance = f (Billing Amount, Property Type, Payment History, Employment Status)

Statistical Model

Probability (Arrears) = f (Billing Amount, Property Type, Payment History, Employment Status)

### 2.2.3 Hypothesis Development

**H1:** Billing amount significantly influences the performance of revenues as well as probability of arrears.

**H2:** Property type significantly affects revenue arrear performance in relation to different property types.

**H3:** Customers with poor payment history are highly likely to become customers in arrears.

**H4:** An economically unstable individual or unemployment raises chances of arrears.

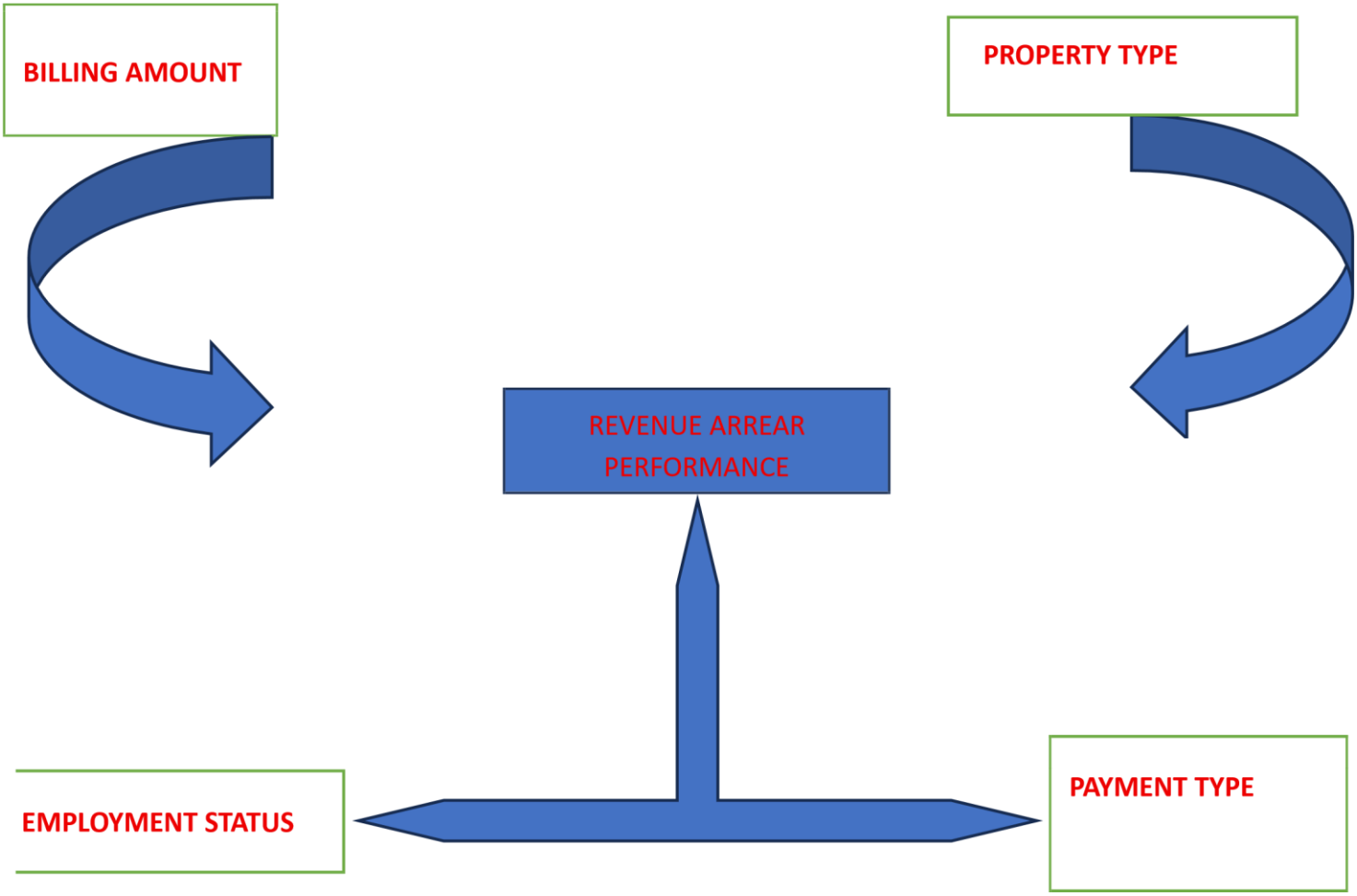
### 2.2.4 Conceptual Contribution of the Study

This conceptual framework integrates both economic capacity determinants such as billing amount and employment status, structural determinant such as property type, and behavioral determinant such as payment history to understand municipal arrears and revenue arrear performance. It provides the theoretical foundation for the use of logistic regression and random forest for classification purposes.

### 2.2.3 Visual Representation of the Conceptual Framework

A visual illustration to show how these elements are interrelated. Below is a conceptual sketch.

**Table 2.1 - Relationship Between the Model and the Real Thing**




### 2.2.4 Implications of the Conceptual Framework

The framework presented above assists in designing the research process as it highlights the way the research should go about collecting and analyzing data. The framework further highlights HOW the research should analyze both the quantitative and qualitative aspects of the research. Through holistically exploring the links latent within the research framework, the research makes recommendations for how Bindura Municipality can best use data science in their municipal revenue management.

### 2.2.5 Conclusion of the Conceptual Framework

The conceptual framework gives a wide perspective that can be employed to conduct the research on the impacts of data science in the municipal revenues management of Bindura. The framework is vital in that it clearly brings out the relationship between the independent, dependent, and

moderating variables that form the basis of the following methodology. This will assist in achieving the objective of improving municipal revenues and governance.

### 2.3 Empirical Literature

This section presents empirical studies in the application of data science methodologies in the analysis of municipal revenues. Focus will be placed on identifying areas of trend, success and improvement in relation to municipal revenues and data science analysis.

#### 2.3.1 Techniques for Revenue Generation

The literature examines different sources of income generation for municipalities ranging from traditional fiscal mechanisms to advanced technological tools. This review of major studies will present the approaches, findings, critiques and applicability to Bindura Municipality according to publication date and author.

- **Bird and Slack (2004)** examine municipal revenue generating frameworks in "Municipal Finance in Really Hard Times." Their research methods include secondary data sources of cross national fiscal statistics, government documents and municipal finance studies. Other approaches include policy comparison, descriptive statistics, and tax theory and reform. The research concludes that property taxation and user fees are vital for financial stability. Recommendations involve increasing fiscal autonomy, valuation system enhancement, broadening the tax base whenever feasible, and streamlining administrative functions to maximize efficiency.

The paper gives an elaborate overview of the approaches and strategies of reforming property taxation, therefore being highly relevant to the research. However, although the strategies are generalizable, they must be customized to fit different contexts. This applies to Bindura Municipality in relation to the strategies of reforming property taxations and administrative reforms. Such initiatives must however depend on local statistics and measures.

- **Mukhongo and Kyalo (2017)** examine the influence of business taxes in the creation of municipal revenue in "Local Government Revenue Generation: Analyzed through Business Taxes". The authors make use of tax records, revenue accounting information, and questionnaires for South African cities. The researchers investigate the impact of business taxation design in determining the level of municipal revenue success with the use of statistical approaches and crosssectional analysis. Moreover, the study makes use of instrument variable technique to address issues of measurement error in business activities. The important finding of the study is the ability

of business taxation policies in generating revenues in the local government without affecting small business formations. Moreover, the research is highly empirical and relies on local administrative data and policies, thus being relevant for implementation. However, the research focuses only on a one-off data collection, hence not allowing for generalizations, especially with regard to small cities with high informal economic activity. Such cities as Bindura could benefit from the customization of taxes and technological improvements in taxation. But the informal economy calls for a different treatment measure.

- **El-Hariri, Ozdemir, and Alghamdi (2021)** in their case study entitled "Predictive Analytics in Local Government Revenue Management" discuss innovative ways of applying modern analytics in revenue management. They consider revenue cycles, taxpayers data, usage statistics, and socioeconomic factors. In testing various predictive models including standard ones as ARIMA and machine learning ones as random forest and gradient boosting (XGBoost), the researchers find ML models, particularly the latter, highly accurate in short-to-medium term cash flow predictions. The researchers advise the use of ML approaches alongside regular updates and sound data structures. However, due to the possibility of overfitting when dealing with limited and unclear data, Bindura Municipality may benefit from using ML approaches once their data infrastructure and capability are fully developed. Until that, simpler predictive approaches would seem more appropriate. Thus, both papers present important lessons on efficient municipal revenue management.

- **Richards and Ruddock (2020)** explore how local government can leverage data analytics in improving tax compliance in "Leveraging Data Analytics for Improved Tax Compliance in Local Government". Tax payment history, demographics, and information on previous compliance interventions form the research data sources. With the use of prediction approaches of logistic regression, decision tree, and ensemble, the researchers conduct AB testing. The result is that targeted interventions perform much better and cost-effectively than general enforcement. The authors propose the use of predictive risk scores in CRM systems and address related ethical concerns. The study is remarkable for its empirical and practical nature and examination of the effectiveness of targeted interventions. Nevertheless, it largely depends on comprehensive administrative data, hence leading to the problem of profile bias and fairness. Such a strategy may increase Bindura municipality's revenues if it develops sufficient data infrastructure and capacity while ensuring proper ethics.

- **Decker, Brown, and Harris (2019)** in their operational case study "Data-Driven Revenue Enhancement in Local Government: Lessons from Los Angeles" examine data integration and anomalies detection in municipal revenue improvement. The researchers use multiple datasets on city finances, different revenues sources, and macroeconomic factors. Through the use of data integration techniques, they conducted exploratory data analysis, detected data anomalies, mapped via GIS, conducted regression analysis, and created dashboards. As a result, they proved the effectiveness of anomaly detection and data integration in improving revenue without tax increases. The strengths of the paper include its empiric, practical, and informative nature, yet it has a weak side in the high capacity of the large city. Nevertheless, Bindura Municipality can implement the strategies described in the paper by utilizing open-source data analytics software and training local employees in cooperation with other institutions.

- **McMillan and Thompson (2020)** focus on the importance of data integration for revenue diversity in "Adopting Analytics: Toronto's Revenue Strategy". Various datasets on municipal finances, revenue sources, and macroeconomic factors serve as the bases for their analysis. The researchers employ data analytics through time series break-down, clustering and what-if simulations. The result is that adopting data analytics makes for financial strength. The research concludes on the necessity of integration of analytics into municipal financial strategies. Although such approach seems obvious and justified, it can be applied by a city with the financial and other resources. Otherwise, Bindura Municipality would find it more difficult to diversify revenue through analytical planning until it manages to establish its own analytical capacity.

- **Becker, Schmid, and Hofmann (2018)** in their article "Challenges of Data Quality in Public Administration: A Systematic Review" examine data quality problems in public administration through the systematic literature review. They sort out data quality challenges by type such as completeness, accuracy, timeliness, and consistency. Moreover, they provide solutions for addressing the identified challenges, including standardized definition of data, ETL processes, validating processes, and employee training. Although the paper draws attention to the problem of data quality as a prerequisite for successful data analytics, it does not present an empirical analysis in a certain locale. As far as Bindura Municipality is concerned, the paper indicates the need for adequate data governance and data validation before engaging in predictive analysis.

- **Lima, Garcia, and Maestre (2019)** examine ways of improving data analytics capacity in local governments in "Enhancing Data Analytics Capacity in Local Governments: Perspectives and Challenges". The authors employ surveys, case studies, and interviews in several municipalities as

their data sources. Qualitative analysis and descriptive statistics enable researchers to identify the factors contributing to organizational readiness for the use of data analytics. According to the findings, proper employee training, recruitment of experienced data analyst, or establishment of analytics unit is crucial. The researchers recommend implementing small-scale pilot programs and scaling the project based on the evaluation results. Although the recommendations are quite clear, the paper highlights the necessity of funding and top management commitment for the realization of the plans. Such initiatives can be started in Bindura Municipality with small-scale pilot programs in collaboration with higher institutions or NGOs.

### 2.3.2 Data Science Applications in Governance

The empirical literature suggests that data science applications are effective for municipal revenue management.

1. **Predictive analytics for Revenue Forecasting:** Recently, El-Hariri et al. (2021) explored the possibilities of predictive analytics in enhancing the revenue forecasting of local governments.

Machine learning approaches helped model historic revenue and forecast high precision in revenue patterns. This contributes to efficient fiscal planning.

2. **Data-driven initiatives for Tax Compliance:** Richards and Ruddock (2020) explored the opportunities provided by data analytics for the improvement of tax compliance. Among their findings was the positive effect of data-driven outreach program on revenues collection.

### 2.3.3 Case Studies of Effective Revenue Management

Several case studies are illustrative of successful data science applications in municipal revenue management:

1. **City of Los Angeles**: The case study by Decker et al., 2019, serves to point out the datadriven strategies set up within the City of Los Angeles toward improvement in business tax revenue. With data analytics and visualization tools, the city improved the speed and accuracy of its tax processing. This contributed significantly to increased collection rates.

2. **Toronto's Municipal Revenue Strategy**: The City of Toronto harnessed the power of advanced data analytics to re-evaluate its revenue streams, according to McMillan and Thompson (2020). Indeed, the results showed how data-informed decision-making yielded greater diversification in the revenue base and enhanced its resistance against fiscal challenges associated with economic downswings.

### 2.3.4 Challenges in Data-Driven Revenue Management

While data science can enhance municipal revenue management, challenges are:

1. **Data Quality Issues**: Becker et al. (2018) find that data quality is one of the main challenges that prevent effective use of data analytics in local government. Poor data quality could result in inaccurate forecasts and inefficient resource allocation, impacting the overall effectiveness of the revenue management strategies.

2. **Capacity Building and Training:** According to Lima et al. (2019), another investment that local governments have to undertake in regard to the analysis involves capacity building and training. Knowledge deficiencies inhibit the use of data science methodologies.

### 2.3.5 Summary of Empirical Findings

According to the empirical literature, the potential for data science to optimize revenue generation in municipalities is increasingly acknowledged. The different cases and studies indicate the benefits obtained through predictive analytics, selective communications, and improved decisionmaking instruments. Data quality and capacity building are some of the challenges that must be dealt with in order to maximize the advantage of using data science.

## 2.4 Literature Synthesis & Research Gap

### 2.4.1 Synthesis of Theoretical and Empirical Literature

From theoretical and empirical literature, a wide range of issues concerning municipal revenue management as well as data science approaches are covered. In particular, the themes from the intersection of the two types of literature include:

1. **Frameworks for Revenue Generation**: In this area, the theoretical literature offers frameworks of how to generate municipal revenues through models of taxation. In the empirical literature, frameworks of the theoretical nature are illustrated by means of case studies that provide insights into best practices that can enhance finances. For example, the framework of Musgrave in terms of guidelines on public finances, as well as the theory of Tiebout, stresses that efficient taxation models should be put in place.

2. **Data Science Applications**:

The findings from the synthesis shows a close alignment between the theoretical perspectives on data science on one side and practice on the other. The existing literature on predictive analytics, machine learning, and similar topics has suggested that data can be used effectively by municipal authorities for improved forecasting of revenues, as well as for efficiently designing tax compliance programs.

**3. Challenges and Institutional Barriers**

Theoretical and empirical literature recognize the difficulties that exist for municipalities, specifically concerning the availability, training, and capacity development. The different theories on institutions and governance, which have been identified, imply that existing structures can be a barrier to the success of applying data management practices. This is supported by empirical literature that has identified instances where the effect of low-quality data has been negative.

### 2.4.2 Research Gap Identity

Even with the existing literature, some research gaps are still evident in relation to municipal finances and the use of data science:

**1. Limited Emphasis on Smaller Municipalities**

Most literature that has been carried out has mainly concentrated on larger cities, while overlooking the existence of other small municipalities such as Bindura, thus providing a chance to examine how data science can assist small municipalities that encounter different challenges with respect to revenues.

**2. Integration of Data Science with Traditional Revenue Models**

Although there is a wealth of literature available on the application of data science, as well as traditional models of public finance, there is a lack of scholarship that combines these two models. There is the potential that research combining models of data science with models of revenues could shed further light on how revenues can be optimized for municipalities.

**3. Lack of Longitudinal**

Most of the empirical research that has been carried out is typically cross-sectional, whereas longitudinal research on the effectiveness of data science application utilization in revenue management would assist in deriving best practices.

**4. Effects of Institutional Change on Revenue Management**

There is very scant research on how changes in the structure of governance, particularly changes in decentralization, affect the management of municipal finances. Such knowledge can increase the efficiency of designing more effective municipal financial strategies.

**5. Evaluative Frames for Data Science Impact**:

The literature appears to lack adequate frameworks within which the effects of data science application practices on municipal finances can be judged. The need for metrics and indicators of performance would ensure that application benefits are measurable in real scenarios.

### 2.4.3 Implications for the current study

The filling of these gaps is paramount in ensuring that research objectives and research methodologies are well-informed. This research intends to bridge the gap within the existing body of knowledge in that it chooses to concentrate on Bindura Municipality, applying data science solutions together with the existing models of revenue generation. The research intends to impact the improvement of revenue management practices by identifying challenges that are peculiar to small municipalities

## 2.5 Summary of Key Findings

1. **Theoretical Insights**: The literature review offered theoretical bases, which includes theories of public finance, the principles of benefit-sharing and ability to pay and models of governance. The models discuss the critical role played by taxation and government spending with regard to the continuity of municipal services. The models discuss the significant aspects of institutional considerations that affect government revenues.

2. **Empirical Evidence**: The empirical literature introduced a set of case studies that illustrated how the application of data science has brought improvements to municipal revenue management. The application of predictive models, visualizations, and automated processing has been illustrated in different cases involving local government. Certain challenges, such as the issue of data quality, have also been recognized, making a negative impact on implementation.

3. **Research Gaps Uncovered**: In a critical analysis, a number of research gaps have been identified, especially in relation to small to medium-sized municipalities, the convergence of traditional means with data-driven solutions, as well as the lack of longitudinal research that investigates the effects of applications of data science.

4. **Development of the Conceptual Framework**: The chapter introduced a conceptual framework which addresses the efficiency of municipal revenues, besides taking into consideration the data science approaches that can moderate it. This helps in specifying the hypotheses on which the research is based, which indicates links between analytics, revenues, and capacity.

### 2.5.1 Connotations for Future Studies

The results and learnings shared within this chapter provide a robust background that aids significantly within the following chapter on methodology. This research, with the aim of filling certain identified gaps, especially concerning the challenges that small municipalities pose, attempts to add to existing knowledge on municipal finances and data science.

The implications that arise from the conceptual framework imply that improved revenue management practices at Bindura Municipality are bound to call for a comprehensive strategy that takes advantage of modern analytical methods, employee capacity, as well as enabling state mechanisms.

### 2.5.2 Getting Ready for the Next Chapter

In the next chapter, there is going to be presented detailed information about the methodology of the research, namely, the research design, methods and analysis, which will help in answering the research questions and testing hypotheses raised within the conceptual framework. As a result, it will be possible to offer solutions for improvement of the revenue base of Bindura Municipality.

## 2.6 Conclusion

In the present chapter, there has been conducted comprehensive literature review on municipal revenues and application of data science approaches. As a result, some important concepts and findings have been highlighted, which prove the necessity of competent revenue management at the municipal level.

# CHAPTER 3: RESEARCH METHODOLOGY

## 3.0 Introduction

This chapter offers a broad picture of the methodology employed in analyzing and forecasting revenue collection from Bindura Municipality based on a data science approach. It covers aspects such as research philosophy, research design, data sources and means of collecting data, data preprocessing, definition of variables and their measuring equations, different types of modeling approaches like Logistic Regression, XGBoost and Random Forest modeling techniques, and finally ethical considerations. Relevant notations and formulae are provided where needed for replicability of the methodologies adopted.

## 3.1 Research Design

This study subscribes to a pragmatist research philosophy emphasizing practical problem solving and the use of adequate methods of analysis to solve problems. The use of this philosophical approach in this study is justified since the goal of this study is to offer solutions based on evidence to improve municipal revenue arrear performance and minimize municipal arrears. This is entirely quantitative in nature as it will only involve the use of secondary data sources from the municipality. This study does not aim at philosophically debating about positivism and interpretivism but instead aims at solving a real-world problem through proper application of statistical and data science techniques. Based on this philosophy, a quantitative analytical approach involving descriptive, diagnostic and predictive analytics will be used in the research process. The research process involves three stages namely, Descriptive analysis which involves the summary of past revenue trends, billing patterns, arrears level and compliance behavior in order to create an organized summary of municipal revenue arrear performance, Diagnostic Analysis where analytic techniques are employed to find patterns and correlation within the administrative data. Patterns and correlation are useful in finding the drivers of municipal arrears and compliance behavior and Predictive Analytics which develops prediction models to predict future revenue trends and account classification based on probability of having arrears in order to help in municipal planning and prioritize on enforcement.

The study employs the use of Logistic Regression, Decision tree, Random Forest and XGBoost models in order to find the main drivers of arrears behavior. The models employ the use of municipal data to determine the direction and relative strength of the predictors on arrears behavior.

Analysis is done at different levels which are, at temporal level where municipal revenue data is disaggregated monthly or quarterly to show revenue trends, at account level where account level administrative data used to build payment behavior model and at geographical level where aggregation at ward level. The timeframe spans across the last years. Preferably, monthly data would be used for better trend identification and improved forecasting accuracy. Longer period helps improve the accuracy of forecasts. The study uses simple statistical models as well as sophisticated machine learning algorithms in order to explain arrears behavior and build accurate prediction models respectively. The Logistic regression model gives insights into the drivers of arrears based on the direction and statistically significant relationship while machine learning models capture complicated patterns within the data.

### 3.2 Data Collection Techniques

This study will make use of only secondary data sources from Bindura Municipality. Secondary Data Sources includes municipal financial records, revenues collected by source e.g., water rate revenues, sanitation revenues, waste management revenues, property rate revenues, billing, payment records, arrears, exemptions and account-level administrative data. Data were extracted using administration procedure which includes extracts from ERP system and financial systems through CSV file download and database export, structured queries in cases where direct access to database was possible, scanning administrative records for digitization in some cases. Anonymizing of data before analysis to remove all personally identifiable information and secure storage protocol adherence.

### 3.4 Data cleaning and Pre-processing

The data set was cleaned prior to modeling. Missing values in important predictor variables were handled through the use of median imputation for continuous variables and mode imputation for categorical variables. Formally, this can be expressed

$$x_i^{\text{imputed}} = \begin{cases} \tilde{x} & \text{if } x_i \text{ is missing and } x \in \mathbb{R} \\ \text{mode}(X) & \text{if } x_i \text{ is missing and } x \in \text{Category} \end{cases} \quad \text{.....(3.1)}$$

where $x$ is the median value of observed values in the feature. Outliers were determined based on the use of domain knowledge and the Interquartile Range (IQR) technique, whereby:

$$\text{Outlier if } x < Q_1 - 1.5 \times IQR \quad \text{or} \quad x > Q_3 + 1.5 \times IQR \quad \text{.....}$$

(3.2)

In an effort to improve the predictive power of the data set, feature engineering was carried out. Features created include the debt to income ratio, loan to value ratio, and payment delay history lagged variables among others, based on the technique proposed by **Khandani, Kim, and Lo (2010)**. Categorical variables such as marital status, occupation, and educational level were coded using one hot encoding, a transformation process which codes each level of the categorical variable in a separate binary indicator. Thus, if a variable has k-levels, it would be transformed into k binary indicator variables.

All continuous variables were scaled using the Min-Max scaling technique, which scales the values of the variables between $[0, 1]$. It takes the form:

$$x_i^{\text{scaled}} = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad \text{..... (3.3)}$$

Finally, the prepared data set was split into three subsets for training (80%) and validation and testing (20%). This was achieved through the use of stratified sampling techniques, which maintains the proportions of classes (i.e. Default and Non-default) of the target variable in all splits **(Bergmeir and Benítez, 2012).**

### 3.5 Variable description

#### 3.5.1 Description of variables - Table 3.1

| Variable Name | Description | Data Type | Sources |
|---|---|---|---|
| Revenue Status | Whether the revenue collected (NO ARREAR=0) or not (ARREAR=1) | Binary | Bindura Municipality |

| | | | |
|---|---|---|---|
| Property Type | Residential or Commercial | Continuous | Bindura Municipality |
| Payment Type | Last payment type (rtgs, ecocash, cash or visa ) | Categorical | Bindura Municipality |
| Payment History | Binary flags of payments made or missed over 24 months | Sequential | Bindura Municipality |
| Billing Amount | Monthly municipal charge | Continuous | Bindura Municipality |
| Employment status | Employed, informal or unemployed | Categorical | ZIMSTATS |

### 3.5.2 Variable justification

1. **Revenue status:** The revenue status indicates whether the ratepayer has paid their municipality payment obligation (0 = NO ARREAR, 1= ARREAR). This is the dependent variable, indicating the outcome I wish to predict.

2. **Property type:** this is to determine if the property is residential or commercial properties. Different types of properties can have different pressure and rates charged (OECD 2019). This will help in predicting which ratepayer, commercial or residential, tends to revenue collections.

3. **Payment type:** is the means used to effect payments. These include the methods available and accessibility **(World Bank, 2021).** Some payment options may be more prone to delayed payments than others.

4. **Billing amount**: **(IMF, 2025)** defines this as the monthly fee charged by the municipality to the ratepayer. The higher fees may pose difficulty to those earning low amounts of incomes.

This can be used in identifying if there are relationships between high charges and the likelihood of revenue collection.

**5. Employment status:** this is whether the ratepayer is employed formally, informally, or unemployed. It shows income stability and is thus a major determinant of whether a person may default.

### 3.6 Models

In this chapter, we discuss models applied in our research. We apply Logistic Regression to classify municipal account arrears status while using Random Forests and XGBoost for easy non-linear modeling and prediction of both classification and revenues. We explain each method in terms of its mathematical basis, estimation process, assumptions, diagnostics, parameters, and interpretation of results.

#### 3.6.1 Logistic regression

In particular, applying Logistic Regression allows estimating the probability of account falling into arrears status. Because the variable of interest is binary and takes only two values, either arrears (coded as 1) or no arrears (coded as 0), logistic regression becomes the best technique for default probability prediction.

Mathematical model formulation

Let denote $Y_i$ for Revenue Status for account $i$,

$Y_i = \{1\ if\ account\ is\ in\ arrears, 0\ if\ account\ is\ not\ in\ arrears$..... (3.4)

The logistic regression model specifies the probability of arrears as

$\pi(X_i = P(Y_i = 1|X_i)$..... (3.5)

The logistic function is

$\pi(X_i) = 1$_____+expexp ($\beta$ ($0\beta+0+\beta 1\beta X11Xi1+i$
$+\cdots\cdots++\beta k\beta XkXki$..... (3.6) This
can compactly written as:

$$\pi(X_i) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)} \quad \text{(3.7)}$$

Equivalently using the inverse logit transformation is:

$$\pi(X_i) = \frac{1}{1+\exp(-X_i\beta)} \quad \text{(3.8)}$$

### 3.6.1.1 HOW LOGISTIC REGRESSION WORKS

Linear regression does not work for binary outcomes since predictions can be outside the [0,1] range and errors aren't uniform. Logistic regression fixes this by modeling the log-odds, or logit of the event happening.

$$\log\left(\frac{i}{1-\pi(X_i)}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} \quad \text{(3.9)}$$

This model is linear in log-odds rather than probability. Plus, the logistic transformation keeps predicted probabilities between 0 and 1.

### 3.6.1.2 Maximum likelihood estimation (MLE).

The likelihood function for $n$ independent observation

$$L(\beta) = \prod^{n}_{\{i=1\}} \pi\,(X_i)^{\{Y_i\}}\,(1-\pi(X_i))^{\{1-Y_i\}} \quad \text{(3.10)}$$

The log-likelihood

$$\ell(\beta) = \sum^{\{n\}}_{\{i=1\}} [Y_i\, log\,(\pi(X_i)) + (1-Y_i) log(1-\pi(X_i))] \quad \text{(3.11)}$$

MLE finds the parameter vector $\beta$ that maximizes

Interpretation

Each coefficient $\beta$ represents the change in the log-odds of arrears for a one unit increase in predictor $X_k$

Exponentiating coefficients gives odds ratio

$$\exp(\beta_k)$$

An odds ratio greater than one implies that increased arrears risk, less than 1 implies reduced risk.

### 3.6.1.3 Logistic regression diagnostics

1. Goodness of fit

- Hosmer-Lemeshow test checks how well the model fits. It splits predicted probabilities into groups and compares observed against expected arrears. When the result is not significant, the fit is good.
- McFadden's pseudo-$R^2$ measures model explanatory power.

$R_{McFadden2} = 1 - \ell\ell((\beta\beta\ _0)$ ..... (3.12)

This compares the fitted model to a null model.

2. Calibration which assesses how well predicted probabilities match observed frequencies.

**Brier Score**

$$BS = -\sum_{i=1}^{n}(\pi(X_i - Y_i)^2 \quad \text{..... (3.13)}$$

Lower values shows better probability calibration

1. Discrimination

• ROC Curve and AUC evaluate how well the model differentiate between arrears and non arrears accounts.

2. Multicollinearity - Variance Inflation Factor (VIF) detects correlation among predictors

$$VIF_k = \frac{1}{1 - R_{k2}} \quad \text{..... (3.14)}$$

High VIF values indicates instability in coefficient estimates.

1. Class Imbalance

If arrears cases are rare, weighted logistic regression or SMOTE resampling may be applied to avoid biased classification

### 3.6.2 Random Forests (RF)

Random Forest (RF) is a supervised ensemble learning approach that builds a forest of decision trees. Its aim is to combine the decisions of each decision tree and thus improve generalization performance. This procedure lowers the variance of the algorithm compared to a single decision tree while keeping the bias minimal.

The training dataset is:

$$D = \{(x_i, y_i)\}^n{}_i = 1 \quad \text{..... (3.15)}$$

- $x_i \in R^p$ is a predictor variables vector of length p
- $y_i \in \{0,1\}$ is the binary response variable (Arrear = 1, No Arrear = 0 • n is the total number of observations

#### 3.6.2.1 Decision Tree Foundation

The decision tree builds a partition of the predictor space recursively by creating nodes and selecting a splitting point which maximizes the improvement of the impurity function. Gini index is a commonly used impurity measure for classification tasks, calculated as follows:

$G(t) = 1 - \sum_{\{k=1\}}^{K} p_k^2$ ..... (3.16)

Where

- $p_k$ is the proportion of observations of class K in node t
- K is the number of classes

The optimal split at note t is selected by maximizing impurity reduction

$\Delta G = G(t) - ({}^{n_L}/_{n_t} G(t_L) + {}^{n_R}/_{n_t} G(t_R))$ ..... (3.17)

Where

- $t_L \; and \; t_R$ are left & right child nodes
- $n_L, n_R$ are observations in child nodes
- $n_t$ is the number of observations in node $t$

### 3.6.2.2 Bagging

Random Forest uses bootstrap aggregation (bagging) to minimize variance.

For each tree $b = 1,2, \ldots \ldots, B$

1. Draw bootstrap sample $D_b$ of size n from $D$
2. Grow a decision tree $T_b(x)$ using $D_b$

Approximately one-third of the observations are not included in the bootstrap sample. They are known as Out-of-Bag (OOB) observations.

### 3.6.2.3 Feature Randomness

Random Forest introduces feature randomness for further decorrelation of trees and variance reduction. At every split use a random subset of size $m$ instead of $p$ predictors. Typically,

$m = \sqrt{p}$…… (3.18)

### 3.6.2.4 Final Prediction Rule

For a new observation $x$ each tree makes its prediction for the class label.

$y_b(x)$…….. (3.19)

The final classification is done by majority voting of predictions of individual trees.

$y_{RF}(x) = mode\{y_1(x), y_2(x), \ldots \ldots, y_B(x)\}$…… (3.20)

Alternatively, if probabilities are considered

$(y = 1|x = \frac{1}{B}\sum_{\{b=1\}}^{B} \widehat{P_b}(y = 1|x)$ $P$ ……(3.21)

with classification criterion, say, probability threshold 0.5.

### 3.6.2.5 Variable Importance

Random Forest estimates predictor importance in different ways.

**(1) Mean Decrease in Impurity (MDI)**

Importance is equal to total Gini impurity decrease for all split points made on that predictor:

$VI_{MDI}(X_j) = \sum_{\{b=1\}}^{B}\sum_{\{t\in T_b:v(t)=X_j\}}\Delta G_{\{t\}}$…….. (3.22)

where:

- v(t) - predictor in the split

- $\Delta G_t -$ Gini impurity decreases at node t

**(2) Permutation Importance**

Evaluates the increase in misclassification error when predictor values are shuffled randomly.

Define:

$Err_{OOB}$ be the out of bag error

$Err_{OOB}{}^{perm(j)}$ be the error after shuffling $X_j$

Importance is computed as follows:

$VI_{perm}(X_j) = Er^{r}{}_{OOB}{}^{perm(j)} - Err_{OOB}$..... (3.23)

The higher increase - the more important is variable. Permutation importance is usually preferred as it has a smaller bias related to having many possible split points for some predictors.

**3.6.2.6 Diagnostics of Random Forest**

**1. OOB Error**

Predictions are made for each observation $i$ based on those trees which have not used $i$ during their training.

The formula for computing out-of-bag error is as follows:

$Err_{OOB} = \frac{1}{} \sum_{\{i=1\}}^{n} I_n (y_{OOB,i \neq y_i})$..... (3.24)

where:

$I(\cdot)$ - is indicator function.

OOB error provides an internal validation procedure without need for an additional test set.

**2. Overfitting Detection**

Overfitting is evaluated by difference between training error and OOB or test error

If

$Err_{train} \ll Err_{OOB}$ ..... (3.25)

The overfitting may occurs. Generally, Random Forest tends to be not overfitted because the number of trees increases.

**4. ROC-AUC Evaluation**

Performance of classification models can be measured using receiver operating characteristic (ROC) curve.

Area Under Curve (AUC) is computed using this formula:

$AUC = \int_0^1 TPR(FPR^{-1}(\mu))d\mu$ ..... (3.26)

Where:

- TPR = True Positive Rate (sensitivity),
- FPR = False Positive Rate.

A higher AUC indicates stronger discriminative power and vice versa

### 3.6.3 XGBoost

XGBoost is a boosting algorithm that sequentially trains decision trees by compensating for the mistakes made by their predecessors. XGBoost has regularization mechanisms that prevent overfitting (Chen and Guestrin, 2016). This algorithm was chosen based on its high prediction accuracy and its efficiency in structured classification problems. XGBoost trains a series of decision trees, where each tree learns from the mistakes of the previous ensemble. This algorithm minimizes the following regularized objective:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad \text{where} \quad \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \qquad \text{..... (3.27)}$$

In this expression, l denotes a differentiable loss function using logistic loss, Ω represents the regularization penalty for the complex model, T indicates the number of leaf nodes per tree, and w refers to the weight coefficients (Chen and Guestrin, 2016). XGBoost performed especially well in analyzing missing payment information, interacting effects between billing cycle, property type, and seasonality, as these factors influence the probability of default by a municipality.

Model Goodness of Fit

$$\text{Brier Score} = (f_t - o_t)^2 \quad \text{..... (3.28)}$$

### 3.7 Model Evaluation and Selection

#### 3.7.1 Evaluating forecasting models

For cross-sectional/classification (logistic regression models): Stratified random K-fold is needed, with temporal integrity preserved (older periods in training set, newer periods in validation)

#### 3.7.2 Metrics of forecast/model performance

Regression/forecasting performance measures:

- $MAE = \frac{1}{n}\Sigma |y_t - y_t|$ MEAN ABSOLUTE ERROR ..... (3.29)
- $RMSE = \sqrt{\frac{1}{n}\Sigma (y_t - \hat{y}_t)^2}$ ROOT MEAN SQUARED ERROR..... (3.30)
- $MAPE = \frac{(100)}{n}\Sigma \left|\frac{y_t - y_t}{y_t}\right|$ MEAN ABSOLUTE PERCENTAGE ERROR..... (3.31)
- $MASE = \frac{MAE}{y_{naive}}$ MEAN ABSOLUTE SCALED ERROR ..... (3.32)

-Mean Absolute Scaled Error (MASE): measures forecast accuracy relative to naive method.

- Prediction interval coverage probability.

#### 3.7.3 Classification performance measures:

- Accuracy, precision, recall, F1-score, AUC, Brier score, calibration plots.

- Consider precision and recall (or AUC) for imbalanced data

### 3.7.4 Model comparison and testing

Test the significance of differences in forecast performance using DM test (Diebold-Mariano) for out-of-sample data

Prefer simplicity: select a parsimonious model with better or about the same performance (Occam's Razor).

Test the performance in different configurations of the data set: feature set, horizon (1 month, 3 months, 6 months), different transformations (nominal or real)

### 3.7.5 Model explainability

Tree-based models: Calculate variable importance, use permutation importance, compute partial dependence plots (PDPs)

Neural Networks: Integrated Gradients, layer-wise relevance propagation, SHAP values.

### 3.7.6 Practical considerations for deployment

- Forecast stability and robustness over time and sensitivity to data gaps.
- Computational cost and frequency of re-training.
- Explainability of model forecasts to municipal officials.
- If two models have comparable accuracy, pick the one providing credible explanations.

## 3.8 Chapter Summary

This chapter described methodology of analyzing and forecasting the municipal revenue of Bindura Municipality using data science techniques. The methods included: research design and mixed-methods, data collection instruments, data cleaning and transformation process, detailed definitions of variables and formulas for constructing variables such as revenue, growth rates, per capita measures, compliance rates; creation of composite SES and service quality indexes. Two types of models were considered: logistic regression Random Forests. The approach to evaluating model performances was specified: rolling cross-validation, metrics such as MAE, RMSE, MAPE, MASE, AUC. Hybrid models and model combinations were considered. Interpretability tools were discussed. Finally, ethical issues of data science research and model deployment were highlighted.

# CHAPTER 4: DATA PRESENTATION, ANALYSIS AND DISCUSSION

## 4.1 Introduction

This chapter will provide an analysis of the empirical results of the Analysis of Municipal Revenues – a Data Science Approach study based on 1499 municipal customer accounts from Bindura Municipality. The chapter will be based on the methodology described in Chapter 3 and will discuss descriptive statistics, preliminary analysis, diagnostics, estimation, validation and results interpretation. The goal is to determine critical factors for successful municipal revenue collection and assess models that predict revenue arrear performance. The primary outcome variable is Revenue Status, which denotes success in collecting revenue (0=NO ARREAR) or defaulting on revenue (1=ARREAR).

## 4.2 Descriptive Statistics

This section offers a statistic overview of the municipal payment dataset used in this study. The dataset contains 1499 customer payments and the summary of key statistics which is billing amount a numerical variable is shown below.

### 4.2.1 DESCRIPTIVE SUMMARY STATISTICS – Table 4.1

| METRIC | REVENUE STATUS | PROPERTY TYPE | PAYMENT TYPE | EMPLOYMENT STATUS | BILLING AMOUNT |
|---|---|---|---|---|---|
| **Mean** | 0.37059 | 1.140588 | 2.507 | 1.892353 | 157.0375 |
| **Standard error** | 0.011191 | 0.008433 | 0.0275 | 0.016822 | 2.362342 |
| **Median** | 0 | 1 | 3 | 2 | 126.74 |
| **Mode** | 0 | 1 | 4 | 2 | 101.83 |
| **Standard deviation** | 0.46141 | 0.347699 | 1.1366 | 0.693576 | 97.40187 |
| **Sample Variance** | 0.212899 | 0.120894 | 1.29188 | 0.481048 | 9487.125 |
| **Kurtosis** | -1.30047 | 2.286805 | -1.40166 | 0.92321 | 4.053557 |

| | | | | | |
|---|---|---|---|---|---|
| **Skewness** | 0.837297 | 2.069811 | -0.00902 | 0.1467 | 2.157832 |
| **Range** | 1 | 1 | 3 | 2 | 523.26 |
| **Minimum** | 0 | 1 | 1 | 1 | 32.36 |
| **Maximum** | 1 | 2 | 4 | 3 | 555.62 |

### 4.2.2 Descriptive Statistics Interpretation

The data sample size is 1499, which is enough for conducting robust statistical analysis and predictions. Revenue Status is the dependent variable; 1 indicates that the account is overdue, while 0 means compliance with payments. Thus, the average value of 0.307 indicates that about 30.7% of accounts are in arrears. Median and mode of 0 confirm that the predominant category among observations is the one associated with compliant accounts. As can be seen, the positive skewness confirms the dominance of the second group. Also, there is a negative kurtosis because the observed distribution is somewhat flat and, therefore, consistent with the one associated with binary outcomes.

It seems that Property Type is another binary categorical variable that distinguishes commercial and residential accounts. According to descriptive measures, the average value is 1.14, while the median and mode equal 1. Therefore, the vast majority of cases correspond to the first group of accounts. The positive skewness equal to 2.07 proves the prevalence of the same category, and the positive kurtosis supports the notion that the distribution is clustered around it. This result is logical since in a typical case, there will be a larger number of residents' accounts in a municipal area. For Payment Type, which takes values between 1 and 4, the average and median are 2.51 and 3 respectively. The close-to-zero skewness proves symmetry around the median. Negative kurtosis proves that there is an almost even distribution of categories, suggesting that none of the payment types dominates over the others. For Employment Status, which takes values from 1 to 3, the skewness is positive, while the kurtosis is negative. Again, this indicates an even balance of categories without any dominating ones.

Finally, regarding the continuous variable called Billing Amount, we can see that the mean and median are 157.04 and 126.74 correspondingly, indicating that there is positive skewness. High skewness and kurtosis, equal to 2.16 and 4.05 correspondingly, prove the above statement. The minimum and maximum values of 32.36 and 555.62 correspondingly, show that there is considerable variance. It also proves a substantial deviation from the mean because the standard

deviation equals 97.40. Generally speaking, this is normal since some accounts will produce bills significantly larger than those from other properties.

Overall, it can be noted that there is some moderate class imbalance in the dependent variable. Also, there is the domination of one category within the categorical variable Property Type. The distributions of Payment and Employment Status are quite balanced. Finally, there is a strong positive skewness in the distribution of Billing Amount.

### 4.2.3 Implications for Modeling

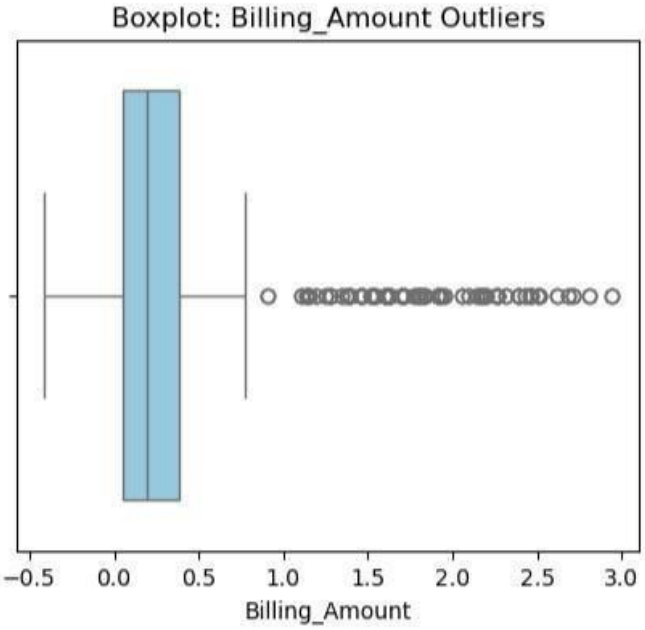
The class imbalance of about 70/30 (non-arrears/arrears) suggests that model accuracy is less important when assessing its effectiveness. Instead, more attention should be paid to sensitivity (arrears recall) measure, precision, F1-score, and area under the receiver operating characteristics curve. Detecting arrears behavior is the key purpose for the model. Moderate dominance of the residential category in Property Type does not represent a particular issue, as it will increase the stability of model performance. Still, it might indicate that structural differences exist between accounts, which makes the variable potentially important for classification. Balanced distribution of categories within Payment and Employment Status makes these variables useful for classification purposes. They may represent independent determinants of arrears behavior. Rightskewed distribution of the continuous variable Billing Amount has certain implications for modeling. While linear models like logistic regression assume linear relationship between predictors and log-odds of the target variable, the extreme values will affect the estimation of coefficients. In contrast, tree-based algorithms like Random Forest and XGBoost are immune to skewness and outliers since they split the data based on thresholds rather than parametric assumptions. Thus, the skewness is potentially valuable for building ensemble-based models.

## 4.3 DIAGNOSTICS TESTS

Prior to estimating the classification models, several diagnostic tests were conducted to assess data suitability, model stability, and statistical validity. They focused on detecting outliers, collinearity, and sample adequacy.

### 4.3.1 Outlier Detection

A boxplot of the Billing Amount feature shows several outliers at the top side of the plot.

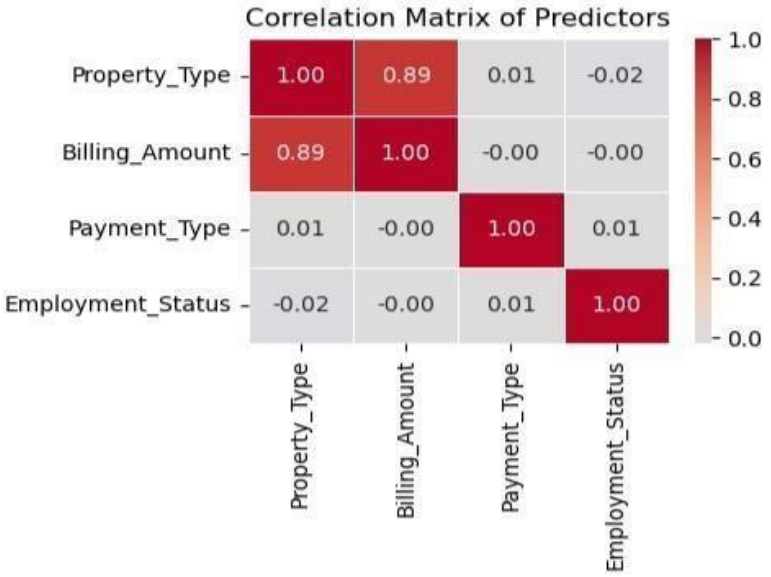



From the Boxplot most billing amounts fall into a narrow interquartile range. There is evidence of some observations above the upper whisker indicating right-skewed distribution. These outliers most likely refer to properties with large consumption volume, accounts of commercial/industrial properties and accounts with accumulated arrears.

Outliers are a natural component of any municipal revenue data due to property type and consumption variation. Given the high robustness of tree-based methods Random Forest and XGBoost outliers pose no problem here. However, for Logistic Regression, skewness can lead to unstable coefficient estimation requiring scaling or transformation. Overall, the presence of outliers reflects normal economic variation without data contamination.

### 4.3.2 Correlation Matrix Analysis

Below is the correlation matrix showing the correlations between predictor features.




The high value of correlation (0.89) means that there is a strong association between billing amounts and property types. It makes sense since there is a difference in tariffs and consumption between residential, commercial, and industrial properties. Other predictor variables Payment Type and Employment Status are almost uncorrelated with other predictors. But a correlation of 0.89 is quite high, it doesn't mean anything yet as correlation does not equal multicollinearity. For this reason, the next step was to run a VIF test.

### 4.3.3 Variance Inflation Factor (VIF) Test

```
===  Variance Inflation Factor VIF ===
             Feature       VIF
0      Property_Type  4.794411
1     Billing_Amount  2.337179
2       Payment_Type  1.000086
3  Employment_Status  2.781375
Rule: VIF < 5 = OK, VIF > 10 = Multicollinearity problem
```

As seen, all VIF values are well below 5. Hence, multicollinearity in this dataset is not that high. Despite the existence of strong correlation, the low VIF values show that there is no distortion in coefficient estimation due to multicollinearity. In other words, all predictor variables were kept in the model.

### 4.3.4 Sample Size Adequacy

```
===  Sample Size Adequacy ===
Total N: 1499
Number of predictors k: 4
Events = 1 count: 461
EPV = Events/k: 115.25
Rule of thumb: EPV >= 10 is adequate. EPV 5-9 = borderline, EPV < 5 = underpowered
```

Given the current EPV of 115.25, the model exceeds the recommended threshold. This means that sample size is sufficient, coefficient estimation is stable, modeling is unlikely to suffer from overfitting, there is a good basis for predictive modeling and conclusively, the sample size is more than adequate for the number of predictors.

### 4.3.5 Summary of Pre-Estimation Tests

Based on the results, we see that there are outliers, but they are real and make sense, there is multicollinearity but not severe enough, the VIF values are acceptable, sample size is more than adequate. Hence, we conclude that the assumptions regarding both logistic regression and machine learning classifiers hold true in this case. This means that the data can be used for estimating the models.

## 4.4 MODEL BUILDING

### 4.4.1 Logistic Regression

#### 4.4.1.1 Model Estimation Results

Logit Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Revenue_Status | No. Observations: | 1499 |
| Model: | Logit | Df Residuals: | 1494 |
| Method: | MLE | Df Model: | 4 |
| Date: | Sat, 06 Jun 2026 | Pseudo R-squ.: | 0.005334 |
| Time: | 20:19:23 | Log-Likelihood: | -920.12 |
| converged: | True | LL-Null: | -925.06 |
| Covariance Type: | nonrobust | LLR p-value: | 0.04271 |

| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -1.8004 | 0.351 | -5.128 | 0.000 | -2.489 | -1.112 |
| Property_Type | 0.8731 | 0.349 | 2.505 | 0.012 | 0.190 | 1.556 |
| Billing_Amount | -0.3842 | 0.200 | -1.925 | 0.054 | -0.775 | 0.007 |
| Payment_Type | -0.0507 | 0.056 | -0.901 | 0.368 | -0.161 | 0.060 |
| Employment_Status | 0.2409 | 0.142 | 1.702 | 0.089 | -0.036 | 0.518 |

A binary logistic regression model with dependent variable Revenue Status was estimated using Maximum Likelihood Estimation (MLE) where the value 1 indicates Arrear and the value 0 indicates Non arrear. The number of observations used in fitting the model was 1,499. The model estimation process converged, implying successful model fit. Model Fit Statistics shows number of observations 1,499, Log-likelihood: -920.12, LL-Null: -925.06, Pseudo R-Square (McFadden) of 0.0053 and the Likelihood Ratio (LR) Test p-value 0.0427. In comparing a model fit against the null model that uses no independent variables, the Likelihood Ratio test is applied. A p-value for the test of $p = 0.0427 < 0.05$ implies that the model is statistically significant. This means that at least one independent variable improves the model fit significantly beyond just having an intercept. However, the very low value of Pseudo-R2 of the model (McFadden's R-Squared = 0.0053) indicates that the independent variables explain very little of the variation in arrears status.

As the coefficients in a logistic regression model are measured on log-odds scale, the coefficient interpretations are carried out based on odds ratios. The model shows us that at 5% level of significant only property type is statistically significant at this level with coefficient of 0.8731 and p-value 0.012. Odds ratio of exp(0.8731) = 2.39 showing that Commercial properties are 2.39 times more likely than residential property type to be in arrears, controlling for other variables. This is a very strong finding of policy relevance. Billing Amount is also marginally significant at 10% level with coefficient of -0.3842 and p-value 0.054. A unit increase in Billing Amount is associated with a decrease in arrears likelihood by about 32%. This implies that higher billed property accounts are less likely to default, perhaps because they have a higher value. However, the marginal significance makes this interpretation subject to some uncertainty. With coefficient of -0.0507 and p-value 0.368, Payment Type is Statistically insignificant and no evidence that payment type matters for determining arrears status.

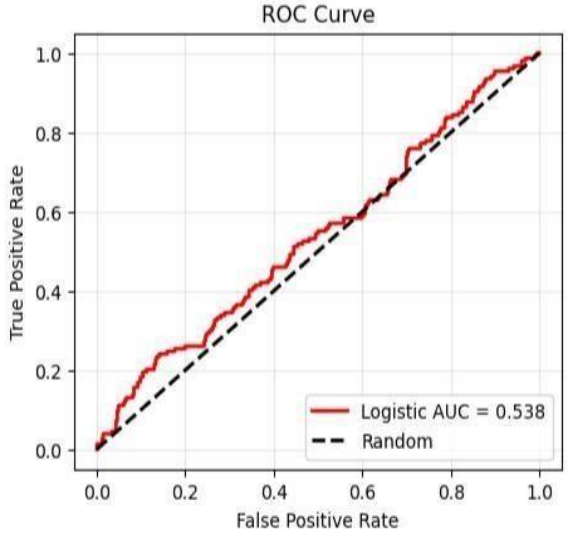
### 4.4.1.2 Classification Performance

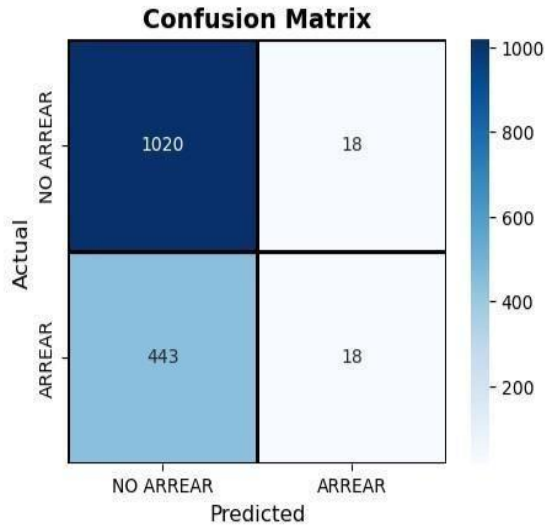
CLASSIFICATION REPORT

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.6972 | 0.9827 | 0.8157 | 1038 |
| 1 | 0.5000 | 0.0390 | 0.0724 | 461 |
| accuracy | | | 0.6925 | 1499 |
| macro avg | 0.5986 | 0.5109 | 0.4441 | 1499 |
| weighted avg | 0.6366 | 0.6925 | 0.5871 | 1499 |

The classifier works exceptionally well at detecting non-arrears accounts. Recall of 98.27% implies that almost all the non-defaulting customers are being detected correctly. A high F1-score (0.8157) further suggests a good performance on this class of no arrear. It justifies the high specificity seen previously. On the other hand, a Recall of 0.039 and F1-Score of 0.0724 implies that the classifier can detect only 3.9% of actual defaulters correctly, of 461 arrears, only 18 cases were correctly classified by the classifier and the F1-score shows an extremely poor predictive accuracy for the arrears. Though the precision is 50%, this is deceptive as the classifier predicts only a handful number of cases under arrears.

### 4.4.1.3 Model Validation




When analyzing the confusion matrix and ROC results provide the following conclusions, the model is highly specific but very poor in sensitivity, cannot identify most of the delinquent

accounts, AUC suggests that its discrimination ability is rather weak and overall prediction is just a little bit above the random chance level. Although the logistic regression is statistically significant according to the likelihood ratio test, being significant statistically does not mean good predictive properties. The logistic regression model seems to be ineffective in predicting the municipal revenues enforcement cases. While it successfully recognizes all the non-defaulter accounts, it fails when recognizing the defaulter accounts, the main one of our interests. With 3.9% sensitivity and AUC of 0.538, the model has rather poor discriminatory abilities for use. The above analysis proves the necessity to consider machine learning algorithms like Random Forests.

### 4.4.2 RANDOM FOREST MODEL

#### 4.4.2.1 Classification Performance

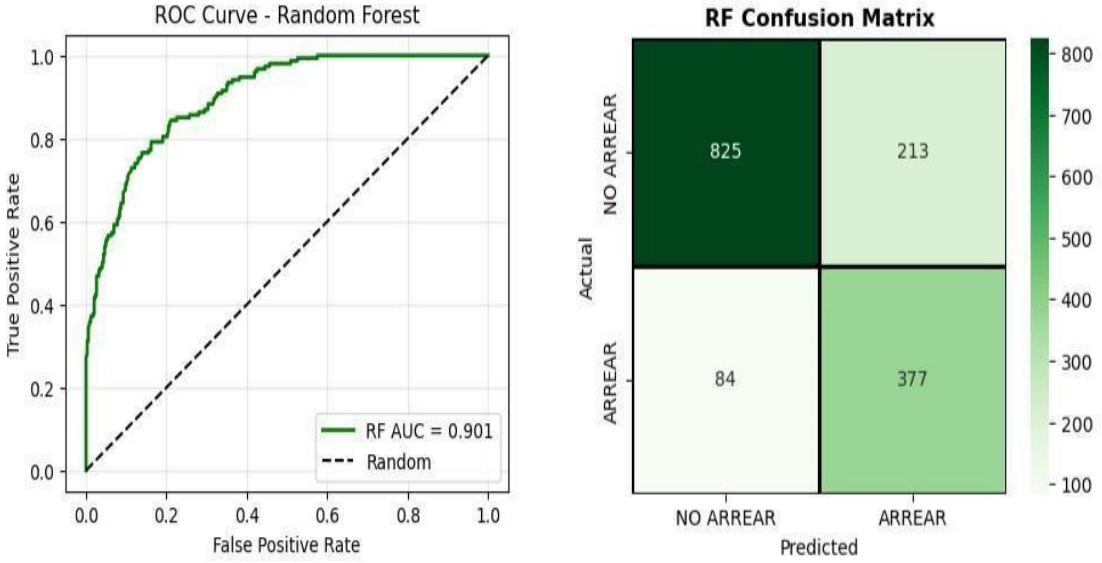
CLASSIFICATION REPORT

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| NO ARREAR | 0.9076 | 0.7948 | 0.8475 | 1038 |
| ARREAR | 0.6390 | 0.8178 | 0.7174 | 461 |
| accuracy | | | 0.8019 | 1499 |
| macro avg | 0.7733 | 0.8063 | 0.7824 | 1499 |
| weighted avg | 0.8250 | 0.8019 | 0.8075 | 1499 |

The Random Forest model was assessed based on precision, recall, F1-score, and accuracy to gauge its predictive power in distinguishing revenue accounts that fall under the category of Arrear or No Arrear. The model provides an overall model accuracy of 80.19%. This implies that roughly 8 accounts were classified correctly out of 10. However, considering class distribution (1038 No Arrear vs 461 Arrear), such measure alone is insufficient to gauge the model, and hence, we will examine class-based measures. In cases where "No Arrear" label is predicted by the model, the prediction is correct 90.8% of times and correctly identifies compliance in 79.5% cases. The F1 score of 0.8475 denotes high precision/recall ratio in case of no-arrears class. It is evident that the

model can distinguish between compliant and non-compliant revenue accounts rather successfully. On the other hand the arrears class is critical from the policy perspective. Recall of 81.78% shows that the model can identify more than four actual defaulters out of five. Model's ability to detect arrears is coupled with sufficient precision: the rate is 63.9%, meaning that the prediction is true almost 2/3 of the times. F1 score of 0.7174 implies a high level of balance between detecting arrears accounts and preventing misclassification. It is apparent that the model can successfully identify problematic accounts, arrears, with satisfactory precision rates. Based on the results obtained, it can be stated that the predictive capacity of the random forest classifier is high in terms of classifying the accounts as being arrears or no arrears.

### 4.4.2.2 Model Validation




Taking into consideration the results obtained from the confusion matrix and ROC analysis together, the following conclusions can be made. Firstly, the model performs quite good overall in terms of accuracy, which amounts to 80%, secondly, the model shows very high sensitivity to arrears detection, which reaches 81.78%, thirdly, the model is characterized by fairly balanced specificity, equaling 79.48%, fourthly, high AUC proves excellent discriminatory power of the classifier and finally, the model provides quite well-balanced trade-off between detection of arrears and false alarms. As opposed to models biased towards one particular category, the Random Forest allows obtaining balanced results for both.

High sensitivity guarantees that most of defaulters will be captured in order to develop further effective targeting approaches. Acceptable precision proves that although some of compliant accounts are likely to be classified as arrears due to the nature of model, the amount of false alarms

is still relatively low. Finally, very high value of AUC proves ranking ability of the model, which allows using it for risk scoring purposes. Overall, it is possible to speak about very high practical usefulness of Random Forest classifier for prediction of arrears at Bindura Municipality.

### 4.4.3 XGBOOST MODEL

#### 4.4.3.1 Classification Performance

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.9565 | 0.9167 | 0.9362 | 312 |
| 1 | 0.8278 | 0.9058 | 0.8651 | 138 |
| accuracy | | | 0.9133 | 450 |
| macro avg | 0.8922 | 0.9112 | 0.9006 | 450 |
| weighted avg | 0.9171 | 0.9133 | 0.9144 | 450 |

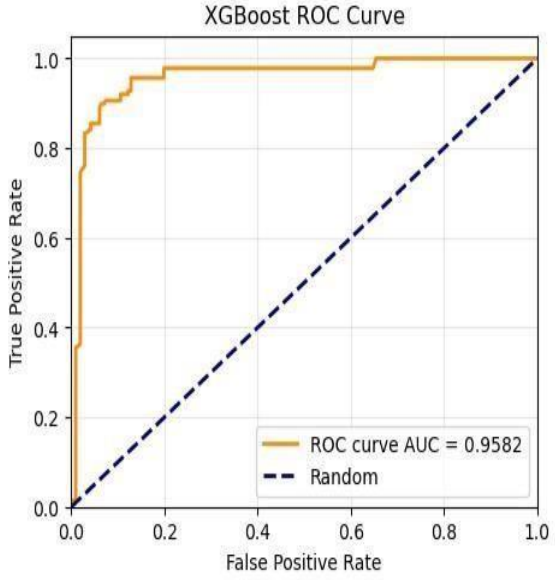
Precision, Recall, F1 score, and overall Accuracy measures were used to validate the efficacy of XGBoost Classifier in categorizing customers as No Arrear and Arrear. The model's overall accuracy is 91.33%. This denotes that more than 90% of the observations can be classified appropriately by the model. The precision measure of 95.65% states that the model predicts non arrears very precisely. The recall value of 0.9167 denotes that 91.67% of the non-arrears cases are identified. The F1 score of 0.9362 demonstrates the model's ability to classify accounts effectively and it shows a high degree of confidence in predicting the category of non-defaulter customers. On Arrears the first priority here is the recall of the arrears since it is more critical. Recall of 90.58% indicates that the model is capable of detecting arrears more than 90% accurately. A precision of 82.78% states that the model correctly labels almost 9 out of 10 cases of arrears. Finally, the F1 score of 86.51% confirms the model's effectiveness in balancing precision and recall. XGBoost is characterized by high global classification accuracy, high recall to detect arrears, high precision, high F1-scores and highly efficient balancing of class-level classification.

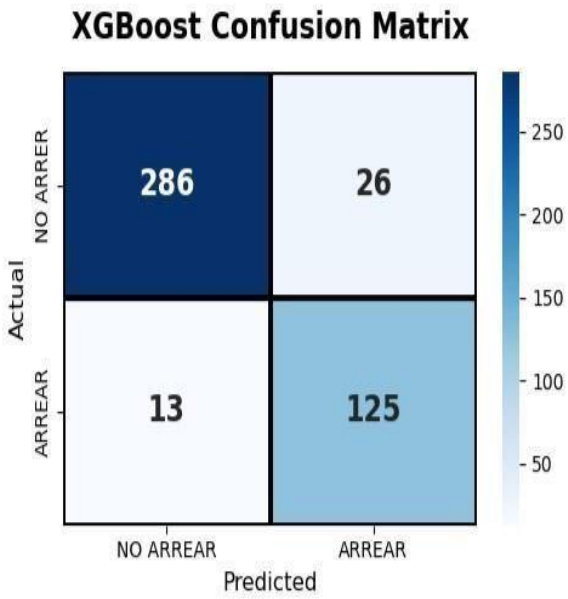
The recall measure ensures that defaulters can be efficiently detected, and high precision avoids unnecessary enforcement actions.

Based on the performance evaluation analysis, it is possible to claim the high quality of the XGBoost predictive model. The model is characterized by a global accuracy higher than 91%, a

high recall of 90.58% for the arrears and a high precision in all classes. Thus, the results of the performance analysis confirm the high efficiency of the predictive tool.

#### 4.4.3.2 Model Validation




The predictive strength of the XGBoost classifier is measured through confusion matrix and ROC analysis. Taking into consideration the confusion matrix and ROC results above, very high overall accuracy of 91.33%, excellent sensitivity 90.58% most defaulters are recognized, excellent specificity 91.67% compliant accounts can be correctly classified, a good precision of 82.78% on classification of arrears accounts and outstanding AUC of 0.9582 showing that the model has excellent discriminatory capabilities.

The XGBoost classifier model shows exceptional predictive strength, classification abilities, and discriminative capabilities. Given that the model is accurate over 91%, with an AUC of 0.9582, the classifier is highly reliable and capable of identifying arrears behavior and making revenue enforcement decisions.

### 4.5 Model Comparison

The selection of the best model framework that could be used to classify cases of arrears was done by assessing the following models, Logistic Regression, Random Forest and XGBoost.

Performance assessment was based on accuracy, sensitivity (arrears recall), specificity, precision, F1-score, and AUC.

**4.5.1 Summary of Model Performance – Table 4.2**

| METRIC | LOGISTIC REGRESSION | RANDOM FOREST | XGBOOST |
|---|---|---|---|
| **Accuracy** | 69.25% | 80.19% | 91.33% |
| **Sensitivity (Arrear)** | 3.9% | 81.78% | 90.58% |
| **Specificity** | 98.3% | 79.48% | 91.67% |
| **Precision (Arrear)** | 50.0% | 63.9% | 82.78% |
| **F1-Score (Arrear)** | 0.072 | 0.717 | 0.865 |
| **AUC** | 0.538 | 0.901 | 0.958 |

The logistic regression model has poor discrimination power with AUC of 0.538, barely performing any better than a random classifier. Random Forest show AUC of 0.901 and has very good discrimination power. XGBoost has great discrimination power AUC of 0.958, practically reaching perfect separation of classes. The gradual improvement of the AUC shows better discrimination between the arrears and non-arrears classes.

This is by far the most significant comparison. On Arrear detection, the logistic model is almost incapable of identifying defaulter clients compared to Random Forest and XGBoost which can identify almost all defaulters. From the point of view of revenue enforcement, this improvement is substantial. Based on the results, we conclude that, there is non-linearity in behavior of the arrears variable, ensemble methods allow us to model it better, XGBoost is the most accurate of all the models due to boosting procedure, which corrects mistakes sequentially and the logistic model

may be inefficient in predicting due to assumptions about linearity of the relation between predictors and log odds. Reasons why XGBoost should be selected as the final model are as follows, XGBoost is the most accurate, has the best arrears recall, highest precision, highest F1Score, highest AUC. The model is not only powerful in terms of identifying defaulters but can also classify them reliably.

It can be stated conclusively that the ensemble machine learning approaches are substantially better than logistic regression at modeling the arrears behavior of taxpayers. In particular, the XGBoost algorithm provides the best overall predictions.

**4.5.2 Revenue Forecasting – Table 4.3**

| | **BILLING AMOUNT** | **RF Prob(ARREAR)** | **XGB Prob(ARREAR)** | **RF-EXPECTED (COLLECTION)** | **XGBEXPECTED COLLECTION** |
|---|---|---|---|---|---|
| 0 | 75.63 | 0.00000 | 0.070088 | 75.63 | 70.33 |
| 1 | 125.5 | 0.8800 | 0.349652 | 15.06 | 81.62 |
| 2 | 119.54 | 0.943333 | 0.706454 | 6.77 | 35.09 |
| 3 | 137.38 | 0.006667 | 0.236719 | 136.46 | 104.86 |
| 4 | 123.59 | 0.000000 | 0.076919 | 123.59 | 114.08 |

**4.5.2.1 Interpretation of Forecast Results**

As mentioned above, the forecast contains predictions of arrears' probabilities produced by the RF and XGB models along with the corresponding expected collection values. All these data items are provided in a single table containing the billing amount, predicted probabilities, and expected collections.

**1. Probability of Arrears**

The two models predict the probabilities of an account going into arrears. For instance, one could note, Observation 1 demonstrates an extremely high probability of arrears according to RF (0.8800) and a moderately high probability according to XGB (0.3497).

Observation 2 demonstrates very high probability of arrears according to both models (RF = 0.9433; XGB = 0.7065), which means that these accounts should be considered risky.

Observations 0 and 4 have very low or even zero probabilities of arrears according to RF and low probabilities according to XGB, which implies the absence of risk.

Thus, the RF model tends to provide probabilities closer to 0 or 1 (extremely low or extremely high) while XGB provides more moderate probabilities.

**2. Interpretation of Expected Collection**

Since expected collection equals the product of the billing amount and the probability of paying the bill (1−P(arrears)), the following conclusions can be made, the higher the probability of arrears, the lower the expected collection and the lower the probability of arrears, the higher the expected collection. For high-risk accounts, observation 2 (probability of arrears = 0.9433) demonstrates extremely low expected collection according to RF (0.0084) as it implies the prediction that no revenue could be collected from this account while expected collection calculated by XGB is a little higher (0.0434) compared to the previous model since its probability of arrears is somewhat lower (0.7065).

For low-risk accounts, observation 4 (probability of arrears = 0.0000) yields high expected collection (0.1737) since RF considers the risk to be close to null whereas expected collection obtained by XGB is slightly lower (0.1603) as the arrears probability is slightly greater (0.0769).

**3. Models' Comparison**

Based on the analysis presented above, there could be found several interesting patterns in predictions of both models, RF produces sharp discrimination between high and low probability accounts Probability predictions of XGB are mostly moderate, therefore producing a smooth curve for expected revenue estimates. Differences in predictions of expected collection values are observed only for those accounts that pose intermediate risk.

Accordingly, it could be concluded that, RF might be better used for classification (high-risk and low-risk accounts identification), but XGB might provide more stable revenue prediction results useful for budgeting and other forecasting goals.

**4. Policy & Revenue Management Considerations**

As one can see, forecasting allows estimating not only probabilities of arrears but also predicting the expected collection for future months based on these probabilities. Therefore, this could help prioritize accounts for which special measures are required, for example, reminders and changing repayment conditions, or, conversely, consider accounts as reliable sources of revenue. By translating classification outcomes into expected revenue estimates, it became possible to estimate the revenue that could be achieved in the upcoming period.

**5. Summary**

Summing up all of the above-mentioned, it could be said that both models were successful in distinguishing high-risk from low-risk accounts. However, while RF tends to produce extreme probability outcomes, XGB yields moderate ones. In terms of translating probabilities into expected revenue collection values, these discrepancies are also visible.

### 4.6 Discussion of Results

This chapter discusses the predictive capabilities of the chosen classification models. Three models, logistic regression, random forest, and XGBoost, are compared in terms of predictive accuracy and discriminative ability, and in terms of how well they are applicable in practice.

As it could be expected, the first model showed statistically significant results in the stage of estimation but poor performance in prediction. Despite having good values of accuracy and specificity, the logistic regression had extremely low sensitivity when it came to detecting arrears accounts. In other words, while being very good at classifying compliant accounts, this method was not able to detect most of the defaulters. In addition to that, the low value of the area under the curve (AUC) is an additional evidence of poor discriminative ability of the method. Overall, it can be stated that a linear functional form is unsuitable in analyzing the arrears' behavior.

The second algorithm random forest showed considerably better predictive performance. As could be seen from results, this technique managed to provide both sensitivity and specificity in a proper proportion, detect all the arrears accounts, and show sufficient precision in doing that. In addition

to that, it provided a very high AUC, which means very good discrimination. Thus, based on the results, it can be concluded that the arrears' behavior depends on non-linear factors and some types of interactions, which the algorithm is able to catch.

Among the studied methods, the best result was shown by the last method, XGBoost. It managed to achieve the highest accuracy, sensitivity, precision, F1-score, and AUC of all the algorithms.

Moreover, the model proved itself very successful in detecting almost all the arrears' accounts and providing high classification reliability. Thus, based on the ROC curve, it has excellent discriminative ability to classify arrears from the compliant accounts. The best performance of the model is a proof of efficiency of gradient boosting and its ability to correct classification mistakes. It appears that the arrears' behavior involves some non-linear interactions, which cannot be captured by classical regression models.

In conclusion, it should be stated that machine learning ensemble methods are highly efficient and effective when it comes to predictive ability. The chapter provides empirical confirmation of that statement.

### 4.7 Chapter Conclusion

This chapter has provided empirical results of the study, analyzing and comparing the performance of three classification models. Accuracy, sensitivity, specificity, precision, F1-score, confusion matrix, ROC curve have been used for that evaluation. The results show that while logistic regression has a limited predictive utility, ensemble machine learning methods increase the performance dramatically. Random forest shows good predictive performance and balance in classifying two types of objects. However, it was XGBoost that has turned out to be the best performing method due to its excellent discriminative power and high accuracy.

Concluding that the arrears' behavior depends on complex and possibly non-linear dependencies, it seems quite clear that XGBoost would be the best approach for this type of classification. This chapter has established a solid base for further discussions.

# CHAPTER 5: DISCUSSION, POLICY IMPLICATIONS AND CONCLUSION

## 5.1 Introduction

This chapter will consider the implications of the findings presented in the previous chapter in detail. For this reason, three classification models were tested during the course of the research – Logistic Regression, Random Forest and XGBoost to determine the most optimal approach for municipal arrears detection. The findings clearly demonstrate that the performance of models varies significantly, especially as concerns the difference between parametric models (like Logistic Regression) and ensemble machine learning techniques, the latter of which proved much more efficient than their counterparts.

## 5.2 Discussion of Key Findings

It should be noted that the main purpose of the study was to design an efficient and reliable predictive model able to recognize accounts likely to go into arrears. As the results reveal, a transition from linear to ensemble machine learning leads to significant improvements in predictive performance.

Even though the logistic regression model showed statistical significance according to the likelihood ratio test, the actual effectiveness of predictions proved very poor, with the model showing a very low sensitivity to detecting cases where customers failed to pay their bills. The algorithm was good at predicting cases where clients paid their bills on time but proved ineffective in recognizing defaulting accounts. The relatively low value of AUC further confirmed the poor predictive performance of the algorithm under consideration. In other words, arrears behavior cannot be reliably predicted by estimating a simple linear relationship between explanatory variables and log-odds.

Random Forest model proved to be a much better fit for prediction of arrears behavior with respect to a number of indicators. First, its sensitivity and specificity scores were relatively high and indicated the good balancing of the predictive model in terms of both true positive and true negative rates. Moreover, high AUC value indicated great discriminatory ability, allowing us to predict the order of accounts according to the risks involved. Thus, we can conclude that arrears behavior is a result of a number of interactions and relations of a non-linear nature and is better understood by tree-based ensemble learning techniques.

XGBoost proved the most effective of all models considered, since it provided the highest scores with respect to accuracy, sensitivity, precision, F1-score and AUC. The algorithm demonstrated high performance both in terms of detecting defaulters and predicting the probability of arrears without increasing the number of false detections. The extremely high AUC score indicated exceptional discriminatory capabilities. In other words, the algorithm demonstrated outstanding performance due to the strength of gradient boosting method. The results prove that arrears behavior can be best analyzed and predicted using advanced machine learning techniques.

### 5.3 Policy Implications

As shown above, the study has certain policy implications that can be implemented.

#### 5.3.1 Targeted Revenue Enforcement

The high sensitivity score of XGBoost implies that the vast majority of defaulting clients can be predicted. This makes it possible to switch from reactive enforcement to proactive approach, since municipalities no longer have to conduct uniform activities aimed at all accounts. Instead, they can focus on high-risk accounts and apply more effective enforcement strategies to ensure clients pay their bills on time.

#### 5.3.2 Risk Based Revenue Management

Thanks to high discriminatory power, it becomes possible to develop a risk-based revenue management framework according to which accounts are classified into groups of low-, medium- and high-risks according to predicted probabilities. This allows municipalities to use different policy instruments depending on the risk level, for example reminders for medium-risk accounts, structured payments plans for high-risk accounts**, r**egular checks and rewards for low-risk accounts.

#### 5.3.3 Evidence-Based Decision Making

Data-driven decision making can help municipalities to make decisions in accordance with empirical evidence and data analysis rather than intuition. The adoption of machine learning tools helps to improve forecasting, budget planning, arrear management and financial sustainability processes in Bindura municipality.

### 5.3.4 Institutional Modernization

Machine learning success calls for the strengthening of institutional and technological infrastructure of municipalities, including the introduction of advanced IT systems for billing and data management.

## 5.4 Theoretical Implications

From the methodological standpoint, the study makes a contribution to emerging discussions about limitations of linear models used in econometrics. Even though logistic regression allows for interpreting coefficients, the model fails to provide accurate predictions of cases involving a large set of independent variables and non-linear dependencies.

The superiority of machine learning algorithms can also be viewed from the theoretical standpoint as confirming the benefits associated with boosting and bagging. Furthermore, the results of the study call for more empirical studies in public finance related to applications of machine learning techniques.

## 5.5 Study Limitations

First, the authors relied entirely on administrative data collected by municipalities. Even though such approach allows for obtaining highly objective and replicable data, it also may be unable to capture important variables influencing arrears behavior. Second, the study focuses on one municipality only, which makes it impossible to generalize the results obtained. Third, even though the predictive performance of machine learning algorithms is extremely good, they can lead to loss of model interpretability compared to linear regression. Fourth, no matter how accurate predictions are, they still may fail due to sudden changes caused by some economic shocks.

## 5.6 Recommendations

1. Apply XGBoost technique to predict municipal arrears.

2. Develop a risk-based revenue management framework using XGBoost as the underlying tool.

3. Design early intervention mechanism targeting medium- and high-risk accounts.

4. Modernize municipal financial infrastructure with regard to data collection and management.

5. Reconduct and validate the model regularly to retain predictive efficiency.

## 5.7 Chapter Conclusion

In conclusion, this chapter has provided discussion of empirical findings obtained during the process of research and confirmed the supremacy of machine learning techniques over logistic regression. As demonstrated above, arrears behavior is governed by complex, non-linear processes that require the use of sophisticated predictive algorithms. In turn, XGBoost proved the most effective and useful of all models analyzed, providing the highest predictive scores. The results of the study confirm that Bindura municipalities can greatly benefit from the use of machine learning algorithms to predict and manage arrears.

## Reference


- Bird, R. M., & Slack, E. (2004). Municipal Finance in Really Hard Times. Public Finance and Management, 4(3), 239–264.
- Mukhongo, L., & Kyalo, P. (2017). Local Government Revenue Generation: Analyzed through Business Taxes. African Journal of Public Affairs, 9(3), 25–37.
- El-Hariri, H., Ozdemir, O., & Alghamdi, A. (2021). Predictive Analytics in Local Government Revenue Management: A Case Study. International Journal of Public Administration, 44(8), 657–670.
- Richards, D., & Ruddock, L. (2020). Leveraging Data Analytics for Improved Tax

Compliance in Local Government. Journal of Accounting and Public Policy, 39(1), 106114.

- Decker, K., Brown, T., & Harris, S. (2019). Data-Driven Revenue Enhancement in Local Government: Lessons from Los Angeles. Urban Affairs Review, 55(3), 743-770.
- McMillan, J., & Thompson, P. (2020). Adopting Analytics: Toronto's Revenue Strategy. Canadian Public Administration, 63(4), 295-312.
- Becker, H., Schmid, L. & Hofmann, E. (2018). Challenges of Data Quality in Public Administration: A Systematic Review. Information Polity, 23(3), 205-223.
- Lima, E. F., Garcia, R., & Maestre, J. C. (2019). Enhancing Data Analytics Capacity in Local Governments: Perspectives and Challenges. Public Management Review, 21(4), 463482. North, D. C. (1990). Institutions, Institutional Change and Economic Performance. Cambridge: Cambridge University Press.
- Musgrave, R. A. (1959). The Theory of Public Finance. New York: McGraw-Hill.
- Tiebout, C. M. (1956). A Pure Theory of Local Expenditures. The Journal of Political Economy, 64(5), 416–424.
- Stiglitz, J. E. (1988). Economics. New York: W. W. Norton & Company.
- Musgrave, R. A., & Musgrave, P. B. (1989). Public Finance in Theory and Practice. New York: McGraw-Hill.
- Govindarajan, V., & Srivastava, A. (2020). Data-driven Decision Making: A Conceptual Framework for Public Sector Organizations. Public Administration Review, 80(2), 305–315
- Chen, J., & Sethi, S. P. (2018). Data Mining and Machine Learning in Public Sector: The State of the Art. Journal of Public Administration Research and Theory, 28(3), 482–495. • Hood, C. (1991). A Public Management for All Seasons? Public Administration, 69(1),
- Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.
- Wooldridge, J. M. (2019). *Introductory econometrics: A modern approach* (7th ed.). Cengage Learning.
- Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.
- Provost, F., & Fawcett, T. (2013). *Data science for business*. O'Reilly Media.
- Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson.
- Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics, 1*(3–4), 323–338.
- Alm, J. (2012). Measuring, explaining, and controlling tax evasion: Lessons from theory, experiments, and field studies. *International Tax and Public Finance, 19*(1), 54–77.
- Slemrod, J., & Yitzhaki, S. (2002). Tax avoidance, evasion, and administration. In A. J. Auerbach & M. Feldstein (Eds.), *Handbook of Public Economics* (Vol. 3, pp. 1423–1470). Elsevier.
- Bahl, R., & Bird, R. (2008). Subnational taxes in developing countries: The way forward. *Public Budgeting & Finance, 28*(4), 1–25.
- Fjeldstad, O.-H., & Heggstad, K. (2012). Local government revenue mobilisation in Anglophone Africa. *ICTD Working Paper No. 7*. International Centre for Tax and Development.
- Smoke, P. (2015). Rethinking decentralization: Assessing challenges to a popular public sector reform. *Public Administration and Development, 35*(2), 97–112.
- Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machinelearning algorithms. *Journal of Banking & Finance, 34*(11), 2767–2787.

- Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series forecasting. *Information Sciences, 191*, 192–213.
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263.
- OECD. (2019). *OECD fiscal decentralization database*. Organisation for Economic Cooperation and Development.