Enhancing Vehicle Licensing Using Machine Learning: A Case Study of the National Road Company

Wallace Chirwa[1], Michael Abednico Chirume[2]

Department of Statistics and Operations Research, National University of Science and Technology, Bulawayo, Zimbabwe

DOI: https://doi.org/10.51584/IJRIAS.2026.11100014

Received: 08 October 2026; Accepted: 13 October 2026; Published: 30 October 2026

ABSTRACT

This paper investigates whether machine learning techniques can enhance vehicle licensing efficiency at the National Road Company (NRC), the body responsible for licensing road vehicles and operating toll gate revenue collection points, whose revenue funds road upkeep and the construction of further toll gates. Using a mixed-methods approach that combined quantitative analysis of historical transaction data for the 2015–2025 study period with qualitative insights from 40 purposively sampled NRC personnel and system users, the research developed and evaluated four machine learning classifiers: Logistic Regression, Decision Tree, Random Forest and XGBoost. The average annual compliance rate across the vehicle fleet was 62.73%, and non-compliance was higher in rural and peri-urban areas (28.5%) than in urban areas (14.7%). On the untouched test set, no model showed meaningful discriminative ability, ROC-AUC values ranged from 0.432 to 0.511, and three of the four models scored below the 0.5 level expected from random ranking. Logistic Regression achieved the highest recall (0.432) and F1-score (0.332), while XGBoost achieved an accuracy of 0.592, a recall of 0.205, an F1-score of 0.240 and a ROC-AUC of 0.475. XGBoost was retained as the framework for further development because it can model non-linear interactions and supports SHAP interpretation, not because it outperformed the other models. SHAP analysis pointed to payment mode, days since last renewal, region type, late renewal history, vehicle class and average renewal lag as the most influential features, although these attributions should be treated as exploratory given the weak model performance. The research contributes to the growing body of knowledge on predictive analytics in public administration and provides practical recommendations for integrating machine learning into NRC compliance management operations, subject to further model improvement and validation.

Keywords: Machine Learning, Vehicle Licensing, Predictive Analytics, Compliance Management, XGBoost

INTRODUCTION

Vehicle licensing plays a cardinal role in promoting road safety, enforcing regulatory compliance and generating revenue for national road funds. The National Road Company oversees vehicle licensing operations; however, the system continues to face significant challenges including compliance gaps, fraud, delays and service accessibility issues. Despite digital upgrades, only approximately 60% of the 1.2 million registered vehicles maintain valid licenses (Mhlanga and Moyo, 2022).

Traditional licensing systems, which often rely on a mix of paper and semi-digital processes, have faced criticism for high administrative costs, corruption opportunities, delays and data loss risks (Smith et al., 2022). As governments strive toward smart governance, the integration of machine learning (ML) and artificial intelligence (AI) into public service delivery has become increasingly prevalent. Systems that can learn from data, spot patterns and make judgments with minimal human intervention represent a significant opportunity for improving regulatory compliance (Turienzo, Cabanelas and Lampon, 2023).

This paper addresses a critical research gap while many studies have examined e-government development

and digitization in Zimbabwe, limited research has explored the application of machine learning to predict vehicle licensing non-compliance from transactional data collected by National Road company. The research aims to investigate how machine learning techniques can enhance the efficiency of vehicle licensing operations, specifically by identifying patterns of non-compliance, developing predictive models and assessing the impact of service delivery channels on compliance rates.

Vehicle Licensing Systems and Non-Compliance

Vehicle licensing systems serve multiple critical functions tracking vehicles, collecting taxes and monitoring compliance with road rules. These systems collect various data points including vehicle class, renewal history and ownership records. However, low and middle-income countries consistently face non-compliance problems such as unlicensed vehicles, expired licenses and unpaid fees. These issues often stem from complex renewal processes, lack of awareness and poor accessibility (Adisa and Eludiora, 2025).

The impact of non-compliance extends beyond revenue loss to include reputational damage and compromised road safety campaigns. In South Africa, automation of licensing regulations has demonstrated reduced non-compliance levels through improved enforcement methods and simplified processes. This highlights the need for data-based approaches that examine non-compliance patterns, particularly in countries like Zimbabwe where internet connectivity and digital literacy vary significantly geographically.

Machine Learning in Public Sector Compliance

Machine learning (ML) has emerged as a transformative tool in public administration, enabling classification, forecasting and anomaly detection in operations (Hotho, 2020). In tax administration, ML has proven effective in detecting fraudulent returns and deviances beyond the capabilities of rule-based systems (OECD, 2023). Governments are increasingly exploring ML applications in social security, environmental compliance and customs operations.

In vehicle regulatory systems, ML can employ prediction models to observe registration and licensing patterns, potentially revealing risky trends such as license renewal delays correlated with certain demographics. However, challenges persist regarding data accuracy, completeness, format consistency and obsolescence. The literature emphasizes significant attention to data preprocessing including imputation, feature selection and normalization to increase efficiency.

Institutional capacity for data management, security and privacy protection represents another critical consideration. The literature suggests that algorithms should incorporate ethical principles in their design to promote non-discriminatory outputs and public confidence in ML applications.

Predictive Modelling for Regulatory Compliance

Predictive modeling using ML algorithms such as logistic regression, random forest, decision trees and gradient boosting has shown promise for achieving regulatory compliance. These models are typically trained on labeled datasets to predict outcomes such as probability of compliance. In the context of vehicle registration, useful predictors include vehicle age, type, renewal history, geographic location and payment trends.

Evidence from vehicle registration systems worldwide indicates that predictive models can support targeted interventions such as reminders and notifications to vehicle owners, improving compliance rates and providing regular income streams for regulators (Zhang et al., 2023). Descriptive analytics can reveal issues and limitations in regulatory systems, while clustering may identify compliance problems concentrated among specific demographics or vehicle characteristics.

However, limitations persist regarding algorithm bias, interpretation challenges and the need for model retraining as behavior patterns evolve. A framework for ethical governance of predictive tools is essential to prevent abuse and protect citizens' privacy. Transparency in how automated scores are derived is particularly

important in contexts where government agencies may not have historically been trusted.

THEORETICAL FRAMEWORKS

This paper draws on three complementary theoretical frameworks:

The Technology Acceptance Model (TAM) (Davis, 1989) posits that two beliefs drive technology adoption. Perceived Usefulness (PU) and Perceived Ease of Use (PEOU). For vehicle licensing, PU relates to whether owners view online platforms as reducing costs and effort, while PEOU concerns how easy the platform is to use. TAM is used in this paper to interpret the varied use of NRC's online licensing platforms.

The Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al., 2003) adds social and contextual factors affecting technology adoption, including performance expectancy, effort expectancy, facilitating conditions and social influence. Facilitating conditions such as internet availability, technical support and agent-assisted service delivery are particularly relevant for understanding compliance.

Compliance Theory (Tyler, 2006) explains that individuals abide by or violate rules based on fairness, legitimacy of the process and ease of compliance. This framework supports the argument that machine learning can aid targeted non-compliance intervention without undermining fairness and accessibility.

METHODOLOGY

Research Design

This paper employed a mixed-methods design combining quantitative and qualitative approaches. The quantitative component involved analysis of National Road Company (NRC) historical vehicle licensing data from 2015–2025, focusing on transaction records including timestamps, payment modes, vehicle status, regional information and compliance records. The study period was 2015–2025, the full period was used for descriptive analysis of compliance trends and as the historical baseline for validation (2015–2024), while model training and testing used a 12-month extract of transactions within this period. The qualitative component comprised semi-structured interviews with NRC employees (IT officers, licensing clerks, managers) and system users to gain contextual insights into system challenges and opportunities.

The research adopted a positivist paradigm, aiming to obtain reliable results through objective data collection and analysis. A case study strategy was employed to investigate NRC's systems and procedures in depth.

Population and Sampling

The target population comprised NRC's vehicle licensing department staff, IT personnel, system developers and customers interacting with the licensing platform. A purposive sample of 40 participants was selected comprising of 15 IT Officers, 10 licensing clerks, 5 senior managers and 10 system users. For the machine learning component, system data samples were extracted from NRC's vehicle licensing database spanning a 12-month period within the 2015–2025 study period.

Data Collection Tools

Data collection integrated three complementary methods which are as below,

Structured Questionnaires: Distributed to NRC staff and customers to collect quantitative data on perceptions of system efficiency, response time and user satisfaction. The response rate was 87.5% (35 complete questionnaires out of 40 administered).

Semi-Structured Interviews: Conducted with 8 key informants including IT managers and system developers to gain qualitative insights into system challenges, data handling processes and ML integration

feasibility.

Documentary Analysis: Examination of NRC's administrative records including licensing transactions, payment records and compliance data.

Data Preparation and Feature Engineering

The raw dataset underwent extensive preprocessing. Removal of duplicate records, correction of inconsistent formats, anonymization of personally identifiable information and handling of missing values.

Feature Engineering derived predictive features including:

- Days since last renewal
- Number of late renewals in the last 3 years
- Average renewal lag days
- Vehicle type and region (urban, peri-urban, rural)
- Preferred payment mode (online, mobile money, agent, cash)
- Time of year (month/quarter)
- Agent synchronization lag

Feature Transformation through Polynomial feature construction: $x_{sq} = x^2$ (days since renewal squared)

- Logarithmic transformation: $x_{log} = \log(1 + x)$
- Interaction feature engineering: $x_{interaction} = x_1 \times x_2$ (payment digital score × region risk score)
- Composite compliance score: Score = $\frac{1}{n}\sum_{i=1}^{n} x_i$
- Cyclical encoding of month: $\sin(2\pi m/12)$, $\cos(2\pi m/12)$

Normalization was done through the Min-Max normalization for fine amounts and Robust Scaling for outlier handling. Feature Selection through SelectKBest method with mutual information reduced the feature space from 49 to 24 variables.

Class Imbalance Handling

SMOTE (Synthetic Minority Oversampling Technique) was applied to generate synthetic instances of the minority class:

$$x_{new} = x_i + \lambda (x_{nn} - x_i)$$

where x_{nn} is the nearest neighbour and $\lambda \in (0, 1)$.

SMOTE addresses class imbalance by generating synthetic observations for the minority class rather than merely duplicating existing records. For each selected minority-class observation, x_i, the method identifies a nearby minority-class neighbour, x_{nn}, and creates a new observation along the line segment connecting the two points. The position of the synthetic observation is determined by λ, a randomly generated value between 0 and 1. This procedure increases the representation of the minority class in the training data and may improve the model's ability to identify non-compliant vehicles. To prevent data leakage and avoid overly optimistic

performance estimates, SMOTE was applied exclusively to the training set after the train–test split, the test set remained unchanged and was used only for final model evaluation.

Machine Learning Models

For the modelling stage, the appropriate machine learning algorithms were chosen by the researcher to predict unlicensed vehicles.

Logistic Regression Model

Logistic regression is one of the most widely used classification techniques and originates in the statistical and probabilistic domain (Hosmer, Lemeshow and Sturdivant, 2019). Because the logistic function models the probability of a binary outcome, it is naturally suited to prediction tasks such as compliance prediction. Its coefficients also indicate the size and direction of the effect of each predictor on the dependent variable. Furthermore, public-sector bodies commonly use logistic regression because it is relatively easy to analyze and interpret (Molnar,2022). Hence, the logistic regression model was examined as a baseline classifier and because it helps NRC policy makers to plan around the factors that motivate vehicle licensing.

Logistic regression model: $\mathbf{P(y = 1|x)} = \frac{\mathbf{1}}{\mathbf{1+exp(-(b_0 + b^T x))}}$

Where $\mathbf{b_0}$ = intercept, $\mathbf{b}$ = coefficient vector and $\mathbf{x}$ = vehicle feature vector.

XGBoost Model

XGBoost was proposed in 2016 by XGBoost to perform complex non-linear interaction of licensing attributes for detecting high risk of unlicensed vehicles. The enhanced boosting model was applied and Log-loss was the chosen cost function, Chen and Guestrin XGBoost performed complex non-linear interaction of licensing attributes for detecting high risk of unlicensed vehicles. The enhanced boosting model was applied and Log-loss was the chosen cost function and it was a fast, scalable, decision tree based gradient boosting algorithm, which is widely used in various fields, especially on classification problems which are often on tabular and structured data. It is an ensemble of weak learners and tries to gradually improve error predictions by following the gradient, thus some main ideas that can be looked at are the tree pruning, regularization and missing value treatment to prevent overfitting and enhance the generalization. Recent works also exploit XGBoost in the government analytics and risk prediction (Zhang et al., 2023), and NRC exploits XGBoost to perform complex non-linear interaction of licensing attributes for detecting high risk of unlicensed vehicles. The enhanced boosting model was applied and Log-loss was the chosen cost function:

$$\boldsymbol{L = -[y\,log(p) + (1 - y)\,log\,(1 - p)]}$$

Where p = predicted probability and y = true compliance label.

Models were trained using an 80/20 stratified train-test split. Class weights were adjusted to handle imbalance.

To prevent data leakage, the stratified 80/20 split was performed first, before any resampling, scaling, feature selection or calibration, so that the test set remained unseen during model development. SMOTE was applied to the training set only and never to the test set. Min-Max normalization, Robust Scaling and SelectKBest feature selection (mutual information, reducing 49 to 24 variables) were fitted on the training data and then applied unchanged to the test data. Engineered features such as days since last renewal and the number of late renewals were computed only from information available before the prediction point. Probability calibration (isotonic regression) was fitted on held-out training data, and all metrics in Table 2 were computed on the untouched test set.

Model selection criteria

The four classifiers were compared using accuracy, precision, recall, F1-score and ROC-AUC. Because non-compliant vehicles form the minority class, recall, F1-score and ROC-AUC were considered more informative than accuracy. XGBoost was retained as the preferred model because of its ability to capture non-linear interactions, its built-in regularization and its compatibility with SHAP interpretation and probability calibration, and not because it was uniformly superior on every metric (see Results).

Model Evaluation Metrics

Confusion matrix is a commonly used evaluation metric. It is a table often used to illustrate how well classification models perform on test datasets with true known values.

Table 1.Confusion Matrix describing the performance of classification models

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

The four parameters in Table 1 are utilized to compute some of the performance measures used to evaluate the model, hence it is important to comprehend their meanings.

- True Positives (TP) - These refer to the number of positive observations predicted by the model that match the actual data, indicating the correct number of non-compliant vehicles predicted by the model and observed in the real data.

- True Negatives (TN) - These refer to the accurate prediction of negative values by the model, based on the actual data. In this scenario, it represents the correct number of compliant vehicles predicted by the model and observed in the real data.

- False Positives (FP) - These represent the positive values that the model has predicted incorrectly, based on actual data. In this case, it refers to the instances where the model predicted non-compliant vehicles that were not observed in the actual data.

- False Negative (FN) - These refer to the negative values that the model has predicted incorrectly, based on the actual data. In this situation, it indicates instances where the model predicted compliant vehicles, but they were non-compliant vehicles in the actual data

Accuracy

The model's accuracy is a metric used to assess its performance. It indicates how well the model can differentiate between compliant and non-compliant cases. However, since this measure does not consider class distribution, it is considered a weak evaluation metric for imbalanced data. The formula for accuracy is provided below:

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

Precision

Precision measures the accuracy of the positive predictions made by the model. Which is defined as follows:

$$\boldsymbol{Precision} = \frac{TP}{TP + FP}$$

Recall

Recall quantifies the proportion of actual positive cases that the model accurately identifies. It underscores the importance of minimizing false negatives, which ensures that as few positive instances as possible are overlooked or misclassified. Mathematically, recall is computed as the ratio of true positives to sum of the true positives and false negatives.

$$\boldsymbol{Recall} = \frac{TP}{TP + FN}$$

The Receiver Operating Characteristic Curve (ROC) - Area Under Curve (AUC)

The area under the curve (AUC) is a metric that assesses the ability of the created model to differentiate between cases. It is the region beneath the Receiver Operating Characteristics curve (ROC), which is a graph that shows the possible distributions of compliant and non-compliant cases. The AUC-ROC curve is utilized to evaluate the performance of a classification task at different thresholds. The ROC is a probability curve, while the AUC measures how well the model can distinguish between classes. A higher AUC indicates a better model. The ROC curve plots sensitivity (true positive rate) versus 1 specificity (false positive rate), with sensitivity on the y-axis and 1 specificity on the x-axis.

Bailey and Law (2006) state that the ROC is a measure used to compare the predicted and actual target values in a classification model. The ROC method is utilized for binary classification where a positive class must be identified. The ROC curve monitors changes in the threshold probability, which is the model's decision point for classification. The default probability threshold for binary classification is 0.5. If the prediction probability is 50% or greater, the model predicts a class, but if the probability is less than 50%, the other class is predicted. The ROC curve represents all the possible combinations of values in the confusion matrix. ROC analysis can determine the probability thresholds that produce the best overall or per-class accuracy. The ROC curve plots True Positive Rates (TPR) versus False Positive Rates (FPR) for the currently selected trained classifier model (FPR).

Model Interpretability Using SHAP

For transparency in enforcement decisions:

$$\emptyset_i = E[f(X) \mid X_i] - E[f(X)]$$

where ϕ_i measures the contribution of feature i

This section outlined the methodology used to investigate the potential of machine learning to improve vehicle licensing efficiency at NRC. It covered research design, data analysis, sampling and ethical considerations.

RESULTS

Demographic Characteristics

The study sample comprised 60% male and 40% female respondents. Age distribution showed 28.5% aged 20–30, 43% aged 31–40, and 28.5% aged 41 or older. Departmental representation included 29% licensing clerks, 14% management, 34% IT/systems and 23% clients. Experience levels showed 26% with 1–5 years, 43% with 6–10 years and 31% with over 10 years patterns of Vehicle Licensing Non-Compliance.

Analysis of NRC's administrative records revealed that 21.8% of 126,450 vehicle licensing transactions between January and December 2025 were either delayed or incomplete. The average annual compliance rate over the 2015–2025 period was 62.73%. The distribution of licensed, unlicensed and deregistered vehicles over this period is presented in Figure 1.



Figure 1. Licensed, Unlicensed and Deregistered Vehicles (2015-2025)

Source: NRC (2025)

Figure 1 shows the numbers of licensed, unlicensed and deregistered vehicles recorded by NRC for each year from 2015 to 2025. The licensed and unlicensed categories show how much of the fleet met its licensing obligations, which gives context for the average annual compliance rate of 62.73% reported above. The deregistered category shows vehicles removed from the register during the study period. Figure 2 complements Figure 1 by expressing annual compliance as a percentage.

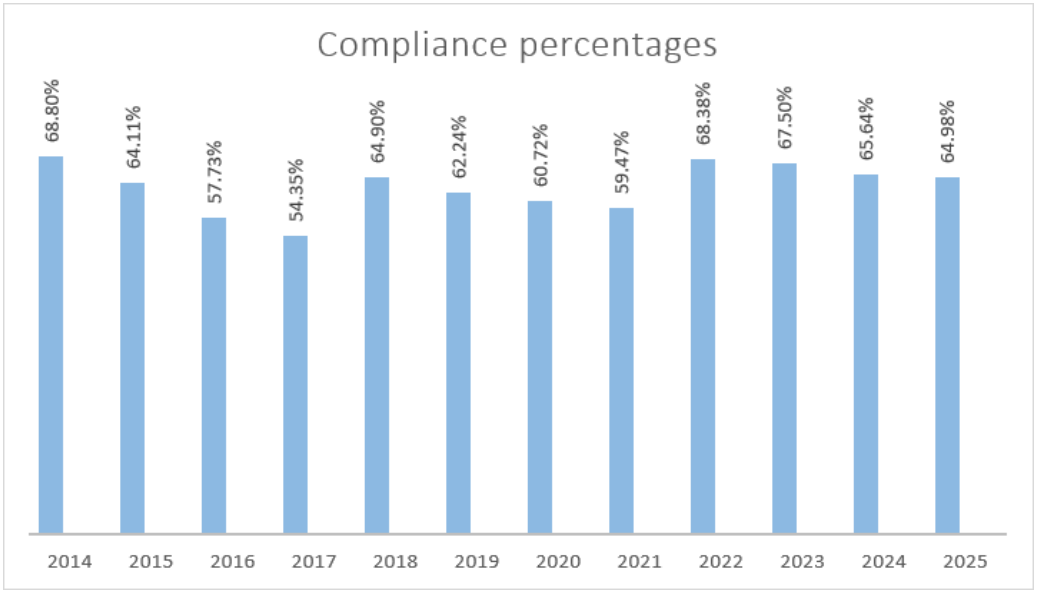


Figure 2. Compliance Percentages Over the Past 10 Years

Source: NRC (2025)

Key findings included non-compliance were more pronounced in rural and peri-urban areas (28.5%) compared to urban areas (14.7%). Licensing activities peaked from January to March, decreasing as the year progressed. Commercial vehicles had significantly higher compliance (89%) compared to privately owned vehicles (73%). Online transactions experienced fewer delays than manual or agent-assisted transactions.

A licensing officer in Bulawayo stated: "Many clients fail to renew their subscriptions promptly, not due to a lack of desire, but rather because the system experiences occasional outages. In such instances, individuals tend to lose motivation and postpone their actions until the following month."

An IT technician underscored data management challenges: "We have duplicate records and incomplete entries from agents. The system flags them, but it is manual to verify. A predictive system that automatically detects anomalies would save us a lot of time."

Machine Learning Model Performance

Four models were evaluated for predicting vehicle licensing non-compliance as reported in Table 2

Table 2. Performance comparison of machine learning models for predicting vehicle licensing non-compliance

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.455	0.269	0.432	0.332	0.432
Decision Tree	0.592	0.331	0.295	0.312	0.511
Random Forest	0.680	0.286	0.014	0.026	0.459
XGBoost	0.592	0.288	0.205	0.240	0.475

XGBoost achieved an accuracy of 0.592, a precision of 0.288, a recall of 0.205, an F1-score of 0.240 and a ROC-AUC of 0.475. Logistic Regression showed the highest recall (0.432) and F1-score (0.332), but the lowest accuracy (0.455) and precision (0.269). The recall of XGBoost was higher than that of Random Forest (0.014) but lower than that of the Decision Tree (0.295) and Logistic Regression (0.432). All ROC-AUC values were below 0.52 (0.432–0.511), indicating limited discriminative ability for every model.

XGBoost was therefore retained as the preferred model because of its higher accuracy than Logistic Regression, its ability to model non-linear interactions and its compatibility with SHAP interpretation, rather than because of superior recall or F1-score, the results should be read as indicative and not as evidence of strong predictive superiority.

Table 3. Confusion Matrix - XGBoost Model

	Actual Positive	Actual Negative
Predicted Positive	30	74
Predicted Negative	116	246

Rows show predicted classes and columns show actual classes. Table 3 presents the confusion matrix for the XGBoost model on the untouched test set (n = 466), which comprised 146 non-compliant and 320 compliant vehicles. The model correctly identified 30 of the 146 non-compliant vehicles (recall = 30/146 = 0.205) and 246 of the 320 compliant vehicles (specificity = 246/320 = 0.769). Of the 104 vehicles flagged as non-

compliant, 30 were truly non-compliant (precision = 30/104 = 0.288), giving an F1-score of 0.240 and an overall accuracy of (30 + 246)/466 = 0.592, consistent with the values reported in Table 2. The 116 false negatives show that most non-compliant vehicles were not detected, while the 74 false positives show that some compliant vehicles would be flagged for unnecessary follow-up. This pattern is consistent with the limited discriminative ability of the model (ROC-AUC = 0.475) noted above.

SHAP analysis identified the most influential predictors as

- Payment mode (online, mobile money, agent, cash)
- Days since last renewal
- Region type (urban, peri-urban, rural)
- Number of late renewals in the last 3 years
- Vehicle class
- Average renewal lag days

Impact of Online Licensing Channels and Agent-Based Service Delivery

Qualitative findings revealed that online licensing portals and mobile payment integrations were introduced to reduce congestion and improve turnaround times. However, server difficulties during peak usage times and data synchronization delays (24-48 hours between agents and NRC's central database) were identified as significant operational vulnerabilities.

A senior manager explained: "Agents help us reach more people, but their systems are not always in sync with ours. Sometimes data from agents uploads days later, which affects enforcement accuracy."

A licensing clerk noted: "Machine learning could help us flag mismatched or incomplete agent entries instantly instead of discovering them weeks later."

Validation and 2025 Compliance Estimate

The model was assessed by comparing historical compliance data (2015–2024) with the compliance rate estimated by the model for 2025. Applying the trained XGBoost model produced an estimated compliance rate of 58.3% for 2025 (based on incomplete data through October 2025), compared to the actual 2024 rate of 63.6%.



Figure 1. Compliance Trends and 2025 Estimate

The 2025 estimate uses the trained XGBoost classification model applied to synthetic feature inputs derived from median values of recent years' data. The model predicts the probability that a vehicle is unlicensed, the compliance rate is 1 − P (unlicensed). Isotonic regression was used to calibrate the predicted probabilities, which improves their reliability but does not guarantee the accuracy of the aggregate estimate. It should be noted that this exercise is predictive classification, not time-series forecasting. The XGBoost model estimates the probability that an individual vehicle is unlicensed from its features, it does not model how compliance evolves over time. The 2025 figure is therefore a scenario-based estimate derived from synthetic, median-based inputs, not an independent out-of-sample forecast, and the difference between 58.3% and the 2024 rate of 63.6% should not be read as a forecast of an actual decline as shown in Figure 1.

DISCUSSION

Patterns of Non-Compliance

The findings indicate that vehicle licensing non-compliance in Zimbabwe is systematic rather than random or uniformly distributed. Higher rates of non-compliance are associated with older vehicles, a history of previous non-compliance, and location in rural or peri-urban areas. These patterns suggest that much of the observed non-compliance is attributable to access barriers and broader systemic constraints rather than deliberate avoidance. The qualitative findings reinforce this interpretation by identifying unreliable connectivity, server downtime, prolonged queuing times, and limitations in database structure as recurring operational challenges.

Machine Learning Model Performance

XGBoost was selected because of its capacity to capture non-linear relationships among licensing attributes and support SHAP-based interpretation however, it did not outperform simpler models across all evaluation metrics. Logistic Regression achieved higher recall and F1-score, while ROC-AUC values below 0.52 for all models indicate limited discriminative performance. Accordingly, the results suggest, rather than conclusively demonstrate, that licensing compliance is shaped by complex interactions. Although the model identified some high-risk vehicles, indicating that proactive compliance management may be feasible in principle, its current performance is insufficient for stand-alone enforcement. In particular, its precision of 0.288 and recall of 0.205 imply that many flagged vehicles would be false positives, potentially resulting in unnecessary follow-up, inefficient use of inspection resources, and diminished trust among compliant vehicle owners, while false negatives would leave some non-compliant vehicles undetected and could reduce revenue collection. These limitations underscore the need for careful threshold selection, human oversight, and further model improvement. Predictive performance may also be affected by data-quality constraints, including duplicate and incomplete agent records, synchronization delays of 24-48 hours and inconsistencies in data from specific years. Any deployment must therefore incorporate appropriate privacy and data-protection safeguards, including the anonymization measures already applied, because the data contain personal and vehicle-ownership information. Model outputs should support rather than replace professional judgement, predictions should be treated as risk indicators, and no penalty or enforcement action should be imposed without review by a licensing officer. SHAP analysis identified renewal history, vehicle class, and payment behaviour as the most influential predictors, consistent with the literature on compliance determinants, while temporal features, including days since last renewal, late renewals, and seasonality, together with geographic features such as region and regional compliance scores, offer potentially actionable insights for targeted interventions.

Service Delivery Channel Impact

Both online and agent-based channels demonstrate potential for improving compliance through increased accessibility and reduced transaction costs. However, their effectiveness is constrained by infrastructure limitations, digital literacy gaps and data synchronization issues. These findings align with the Technology Acceptance Model (TAM) and UTAUT frameworks, confirming that perceived ease of use and facilitating conditions significantly influence technology adoption and, consequently, compliance behavior.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue X October 2026

CONCLUSIONS

Vehicle licensing non-compliance in Zimbabwe appears to be driven primarily by structural constraints rather than individual behaviour, indicating that effective responses should combine enforcement with broader systemic reforms. Machine learning has potential as a decision-support tool for proactive compliance management, as demonstrated by the XGBoost model developed in this study, however, its limited discriminative performance must be improved and rigorously validated before operational deployment. Although XGBoost's capacity to model non-linear interactions indicates that vehicle licensing compliance is shaped by complex relationships among multiple factors, its overall performance was modest and did not consistently exceed that of simpler models across all evaluation metrics. Online and agent-based service delivery channels can also improve compliance by expanding access and reducing transaction barriers, but their effectiveness depends on reliable infrastructure, adequate digital literacy, and effective data integration across service platforms.

Recommendations

Integration of Machine Learning into NRC Operations

- Pilot the XGBoost model, with further validation and human review, within NRC's licensing and compliance management system
- Update predictive models quarterly to address behavioral changes and data drift
- Develop an operational dashboard for staff to access predictions and prioritize inspections

Improvement of Data Infrastructure and Data Quality

- Commit to an integrated data management solution supporting real-time analytics
- Reduce incomplete and duplicate records through standardized data entry protocols
- Implement a clear data governance framework

Expansion of Online Licensing Platforms

- Invest in user-friendly interfaces, mobile applications and localization
- Conduct public awareness campaigns targeting rural and peri-urban populations
- Enhance server capacity to handle peak usage periods

Strengthening Agent-Based Service Delivery

- View agent networks as strategic complements to online channels
- Provide training, equipment and real-time data synchronization tools
- Establish accountability frameworks to protect agent network integrity

Capacity Building and Staff Training

- Develop training programs in data analytics, machine learning, ethical AI and cybersecurity
- Build institutional capacity to sustain intelligent licensing systems

Policy and Regulatory Support

- Establish enabling policy frameworks supporting digital transformation and responsible AI deployment
- Address data privacy rights, algorithmic transparency and accountability mechanisms

Limitations

This paper has several limitations including data quality issues in specific years, the case study design limits generalizability, the sample size (40 participants) may not fully represent all stakeholders and computing power constraints limited exploration of more complex models such as deep learning. In addition, the discriminative performance of all models was limited (ROC-AUC below 0.52), and the 2025 estimate relied on synthetic inputs.

Areas for Future Research

Future research should investigate novel AI techniques including deep learning, neural networks and reinforcement learning for compliance prediction. Integration of ML with Automated Number Plate Recognition (ANPR) systems, blockchain vehicle registries and IoT devices for compliance monitoring represent promising directions. Comparative studies across countries and regulatory regimes would help identify institutional and technical settings that maximize ML-based compliance solutions.

REFERENCES

1. Adarmouch, L., Sebbani, M., Elyacoubi, A., & Dahhou, M. (2020). Scientific research and ethics: Foundations and principles in research involving human subjects. International Journal of Ethics Education, 5(2), 145-156.
2. Adisa, O., & Eludiora, S. (2025). Digital transformation and vehicle licensing compliance in developing economies. African Journal of Public Administration, 14(1), 34-49.
3. Ayo, C., Oni, A., Adewoye, O., & Eweoya, I. (2022). E-government adoption and intermediary service delivery in Africa.
4. Government Information Quarterly, 39(3), 1-12.
5. Badi, I., Bouraima, M., & Muhammad, L. (2021). Intelligent transportation systems and public sector efficiency in developing countries. Transport Policy Review, 18(2), 77-90.
6. Bosire, M., & Maingi, J. (2022). Digital transformation challenges in sub-Saharan African transport systems. African Journal of Information Systems, 14(3), 55-69.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).
8. Chingovo, T. (2021). Digital Transformation in Zimbabwean Public Institutions: Opportunities and Challenges. Harare: University of Zimbabwe Press.
9. Creswell, J. W. (2014). Research Design: Qualitative, Quantitative and Mixed Methods Approaches (4th ed.). Thousand Oaks, CA: Sage Publications.
10. Davis, F. D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. MIS Quarterly, 13(3), 319-340.
11. Denhere, V., Moyo, P., & Sibanda, T. (2024). Artificial intelligence and regulatory compliance systems in Zimbabwe. Zimbabwe Journal of Information Technology, 9(1), 22-38.
12. Hastie, T., Tibshirani, R., & Friedman, J. (2017). The Elements of Statistical Learning (2nd ed.). New York: Springer.
13. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2019). Applied Logistic Regression (3rd ed.). Hoboken: Wiley.
14. Inyangala, B., Omwansa, T., & Mwangi, W. (2021). Smart mobility and intelligent transport systems in Africa. African Transport Review, 10(2), 41-57.
15. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning (2nd

ed.). New York: Springer.
16. Khan, R., Karim, S., & Syed, H. (2022). ICT infrastructure and digital governance in developing economies. Journal of Information Systems Development, 17(2), 88-103.
17. Kuhn, M., & Johnson, K. (2022). Feature Engineering and Selection: A Practical Approach for Predictive Models. Boca Raton: CRC Press.
18. Mhlanga, T., & Moyo, P. (2022). Vehicle licensing compliance challenges in Zimbabwe. Zimbabwe Economic Review, 12(2), 88-101.
19. Molnar, C. (2022). Interpretable Machine Learning (2nd ed.). Munich: Leanpub.
20. Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2019). Handbook on Constructing Composite Indicators. Paris: OECD Statistics Working Papers.
21. OECD. (2023). Artificial Intelligence in Tax Administration and Public Compliance Systems. Paris: OECD Publishing.
22. Saunders, M., Lewis, P., & Thornhill, A. (2019). Research Methods for Business Students (8th ed.). Harlow: Pearson Education.
23. Smith, J., Brown, T., & Wilson, R. (2022). Challenges in traditional vehicle licensing systems. Journal of Transportation Administration, 9(2), 55-70.
24. Turienzo, A., Cabanelas, P., & Lampon, J. (2023). Artificial intelligence and machine learning in public administration. Government Information Quarterly, 40(2), 1-13.
25. Tyler, T. (2006). Why People Obey the Law. Princeton: Princeton University Press.
26. Venkatesh, V., Morris, M., Davis, G., & Davis, F. (2003). User acceptance of information technology: Toward a unified view.
27. MIS Quarterly, 27(3), 425-478.
28. World Bank. (2021). Digital Government and Service Delivery. Washington, DC: World Bank Publications.
29. Zhang, Y., Li, J., & Chen, X. (2023). Automated number plate recognition and intelligent vehicle surveillance systems. International Journal of Intelligent Transport Systems, 1(1), 52-68.
30. ZINARA. (2025). Vehicle Licensing Statistics: Mid-Year Report 2025. Harare: Zimbabwe National Road Administration.