An Adaptive Telematics Framework for Usage-Based Insurance in Kenya

By

Kiruki Ruth Wambui

54667

Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Data Science at Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

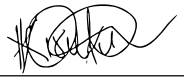
Nairobi, Kenya

June, 2025

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Kiruki Ruth Wambui



29/05/2025

Approval

The dissertation of Kiruki Ruth Wambui was reviewed and approved by the following:

Professor Dr. Bernard Manderick

The Free University of Brussels.

Dr. Godfrey Madigu

Dean, Strathmore Institute of Mathematical Sciences, Strathmore University.

Prof. Bernard Shibwabo

Director of Graduate Studies, Strathmore University.

Abstract

This study develops a telematics-based risk classification model for usage-based insurance (UBI) in Kenya's fleet management sector, addressing critical limitations of traditional models through adaptive machine learning techniques. The research employs ADASYN (Adaptive Synthetic Sampling) to resolve class imbalance. This approach enables more effective learning of risky driving patterns compared to conventional sampling methods.

The study processes telemetry data—including acceleration, deceleration, RPM, and speed—from Easy Coach Limited vehicles collected between January 2022 and February 2024. After minority class augmentation via ADASYN, several machine learning models were evaluated, including logistic regression, random forests, and gradient boosting. Among these, XGBoost demonstrated optimal performance, achieving a classification accuracy of 99% and an area under the ROC curve (AUC) of 0.98. Notably, the model achieved a precision of 93% and recall of 97% for the minority (risky) class. ADASYN yielded a 27% improvement in the minority class F1-score compared to SMOTE, highlighting its effectiveness in balancing highly skewed datasets.

The model was deployed on an Amazon EC2 instance within a secured Virtual Private Cloud (VPC). The instance hosts a Gradio-based interface, which generates a URL endpoint that can be embedded in external web applications. This allows seamless, real-time interaction with the model. Operational monitoring was implemented using Amazon CloudWatch for system health tracking and failure alerts. Model explainability was supported through SHAP (Shapley Additive

Explanations), providing transparency into feature importance. Key applications of the model include precision-targeted premium adjustments based on risk scores, temporal risk analysis and operational optimization based on RPM and braking behaviors.

This work contributes empirical evidence supporting a deployable framework for insurance technology in Kenya. Furthermore, it proposes policy guidelines for the insurance industry to support UBI adoption. Future research directions include applying federated learning to enhance model generalizability across East African fleets while addressing data privacy concerns.

Table of Contents

List of Figures

List of Tables

Abbreviations and Acronyms

ADASYN – Adaptive Synthetic Sampling

AUC – Area Under the Curve

AWS – Amazon Web Services

EC2 – Elastic Cloud Compute

EDA – Exploratory Data Analysis

FN – False Negative

FP – False Positive

GDPR – General Data Protection Regulation

GPS – Global Positioning System

JSON – JavaScript Object Notation

LIDAR - Light Detection and Ranging

MHYD – Manage How You Drive

ML – Machine Learning

MPG – Mile Per Gallon

ONNX – Open Neural Network Exchange

PAYD – Pay-As-You-Drive

PHYD – Pay-How-You Drive

PII – Personal Identifiable Information

PSV – Public Service Vehicle

TP – True Positive

TN – True Negative

RPM – Rotations Per Minute

SHAP – Shapley Additive Explanations

SMOTE – Synthetic Minority Over-sampling Technique

SQL – Structured Query Language

UBI – Usage-Based Insurance

URL – Uniform Resource Locator

VPC – Virtual Private Cloud

XGBoost – Extreme Gradient Boosting

Acknowledgements

I would like to express my sincere gratitude to my advisor, Professor Dr. Bernard Manderick, for his invaluable guidance throughout this research. Special thanks to Easy Coach Limited for their unwavering support and collaboration throughout this research. Their willingness to share operational data was invaluable, providing the foundation upon which this study was built. I deeply appreciate their continued commitment to innovation and their pioneering efforts in adopting telematics to advance safety and efficiency in the public transport sector.

I am profoundly grateful to my parents, whose steadfast support, encouragement, and belief in my academic journey made this milestone possible. Their sacrifices have been my greatest source of strength.

I also extend my heartfelt appreciation to my colleagues for their insightful feedback, solidarity, and shared pursuit of excellence. Their enthusiasm for data science and machine learning created a stimulating environment that constantly challenged and inspired me.

I am deeply grateful to both the Strathmore Institute of Mathematical Sciences, iLabAfrica and Strathmore University for their comprehensive institutional support, including access to critical research facilities and academic resources, which proved invaluable throughout this research endeavor.

1.0 Introduction

Insurance functions as a mechanism for transferring risk, providing the insured party with partial or complete financial compensation for losses or damages resulting from events outside of their control. Insurance companies assume these risks in exchange for a predetermined amount of money paid by the insured in advance, called a premium (Macedo, 2009). To earn a reasonable profit, insurance companies must manage these risks appropriately by charging a fair premium for the assumed risks. The premium should not be too high, causing customers to lose interest, nor too low, jeopardizing insurer profitability. A fundamental principle in this risk management process is known as "risk classification". Insurance companies group individuals with similar traits into a large pool and assume that they have the same expected costs (Finger, 1988).

An actuary then calculates a price for the group, with each member receiving the same premium, regardless of individual risk characteristics. The traits used by insurance companies to group people together are referred to as "rating variables" (Finger, 1988). In the auto insurance industry, insurers frequently consider both driving-related and non-driving-related rating variables when pricing their policies.

Companies that have a large number of vehicles under their management have led to the use of telematics. The term "telematics" refers to the technology and system that gathers and transmits data from vehicles, assets, or equipment to a distant location for uses such as tracking, monitoring, and analysis. Typically, real-time data collection methods include sensors, GPS (Global Positioning System), and computers on board to collect position, speed, fuel economy, and vehicle diagnostics. In order to improve operational efficiency and safety through data-driven insights,

telematics is widely utilized in the automotive industry for fleet management, insurance, and navigation systems (Alamir et al., 2020).

Easy Coach Limited has been able to achieve this by enrolling its buses in a fleet management system that provides them with oversight. This system has allowed them to keep track of their performance on the road and ensured the safety of their fleet and passengers. Through this, the company has been able to maintain leadership in this sector. The management of commercial fleets involves several crucial aspects, one of which is motor insurance. Motor insurance is an essential component of fleet management as it protects vehicles and drivers against various risks, including accidents, theft, and damage. The cost of motor insurance can be a significant expense for fleet managers, especially those who manage large fleets, making it essential to have an effective pricing model.

The pricing of motor insurance is a complex process that involves several factors, including the age and type of vehicle, driver experience, driving patterns, and claim history. Traditionally, the pricing of motor insurance has been based on a variety of demographic factors, such as the age and gender of the driver, the location of the vehicle, and the type of vehicle. However, this approach does not accurately reflect the unique risk profile of a vehicle.

This dissertation aims to develop a model for motor insurance for fleet-managed vehicles that takes into account the specific characteristics of these vehicles and their drivers. The proposed risk assessment model will provide a more accurate and fair pricing structure for insurance, considering the unique risk profile of each fleet.

The framework of this dissertation will begin with a study of the body of knowledge about automobile insurance models. The literature review will highlight critical elements that affect insurance

costs and give a thorough overview of the state of the field as a whole. This will dive into the existing methods and algorithms that are used in the industry.

The methodology section will explore the research strategy and techniques utilized to create the suggested model. The data sources and analytical strategies utilized to create the model will be described in this section. The study findings, including the creation of the suggested model, will be presented in the findings section. The results will be presented succinctly and clearly and suitable visual aids such as graphs and charts will be used for the presentation of the data.

A critical evaluation of the findings and their implications for the cost of auto insurance will be provided in the discussion section. This dissertation will be of particular interest to fleet managers, insurance companies, and policy makers involved in the regulation of this field.

1.1 Research Objectives and Questions

1.1.1 Research Objectives

i) Identify the factors that influence a usage-based insurance model.

ii) Assess the laws and regulatory framework that surround the motor insurance industry.

iii) Create a study of the benefits of using a usage-based insurance model vis a vis the current model.

iv) Develop a usage-based insurance model for the Easy Coach fleet in Kenya.

1.1.2 Research Questions

i) How does a Usage-Based Insurance model compare to traditional insurance models in terms of cost and effectiveness?

ii) What factors affect driver behavior in a Usage-Based Insurance model?

iii) What impact does driver behavior have on the cost of insurance in a Usage-Based Insurance model?

iv) How effective is a Usage-Based Insurance model in reducing risky driving behavior?

v) How do different incentives and rewards affect driver participation and behavior in a Usage-Based Insurance model?

vi) How does the age and experience of the driver impact the effectiveness of a use-based insurance model?

vii) What is the impact of external factors such as weather, traffic, and road conditions on driver behavior in a Usage-Based Insurance model?

1.2 Scope and Limitations

1.2.1 Scope of the Study

The scope of the study includes the acquisition of an appropriate dataset. The dataset relating to telematics was obtained from Easy Coach Limited, which operates in Kenya and has developed a full-fledged management system for its fleet. In addition, gaining insight into insurance pricing

models and regulations from a local insurer. The data collection exercise was conducted at the participants' workplaces or by telephone, accommodating their availability and locations.

1.2.2 Limitations

Machine learning (ML), although it has the potential to bring about significant changes in different areas, faces certain restrictions that hinder its effectiveness and ethical use. ML results are essentially influenced by the quality and quantity of training data. If the data is biased or of poor quality, it might result in inaccurate or unfair models. This highlights the importance of carefully curating a robust dataset.

i) Data availability

Data scarcity is a crucial constraint that significantly impacts the advancement and implementation of ML models. Ample quantities of superior data is needed for successful training of several machine learning algorithms, particularly those that incorporate deep learning techniques. Constructing resilient and precise ML models becomes arduous when faced with situations where pertinent data is few, antiquated, or difficult to acquire.

Insufficient data availability can lead to poor model generalization and heightened vulnerability to over-fitting. Over-fitting arises when a model acquires an excessive understanding of the training data, yet struggles to apply this knowledge to novel, unobserved data. This constraint is especially evident in specialized domains or emerging sectors where there is a scarcity or absence of labeled datasets. A labeled dataset is a compilation of data in which each individual data point is assigned one or more labels that indicate certain classifications. These labels are used in supervised learning tasks within the field of machine learning (Nasteski, 2017). Moreover, in specific sectors where

privacy considerations limit data sharing, acquiring a sufficient amount of varied data for thorough model training presents a significant challenge.

ii) Data quality

The application of ML models is significantly hindered by the constraint of data quality. The efficacy and precision of ML algorithms are greatly dependent on the quality of the input data. Model performance and prediction reliability can be compromised by inconsistent, incomplete, or biased data. Problems such as the absence of data, mistakes, or uneven distribution in the dataset can induce biases and alter the learning process of ML models, potentially resulting in erroneous outcomes.

Data quality assurance encompasses thorough methods for data cleansing, preparation, and validation. This constraint is further intensified by the requirement for substantial amounts of superior labeled data for supervised learning. Acquiring such data, particularly for specific or specialized fields, might be difficult in certain cases. Data quality challenges are further intensified in dynamic systems characterized by fluctuating data patterns. To address these problems, it is necessary to make a focused and coordinated effort in implementing data governance, quality assurance processes, and ongoing monitoring to improve the dependability and accuracy of the data used to train and deploy ML models.

iii) Time frame

The time frame poses a significant constraint in the context of ML models. Creating and educating resilient ML models frequently requires significant time commitments. The problem within this period is complex, encompassing the duration required for gathering and organizing data of supe-

rior quality, going through multiple cycles of model development and optimization, and ultimately implementing the model for practical application. The prolonged duration can impede the prompt execution of ML solutions in rapidly developing domains or situations necessitating immediate actions.

Furthermore, the temporal dimension is crucial when considering the perishability of models. With the progression of technology, the models might become obsolete, requiring ongoing updates and adjustments. The rapid rate at which data patterns evolve and the constantly changing nature of real-world situations also contribute to the issue of managing time frames. Achieving a harmonious equilibrium between the requirement for precise and skillfully developed models and the urgency to quickly implement them continues to be a continuous challenge. The statement emphasizes the importance of well-organized processes, effective model structures, and flexible approaches to overcome time limitations in the field of ML development and implementation.

iv) Data Complexity;

The complexity of the data presents a substantial obstacle in the creation and implementation of ML models. The difficulty stems from the heterogeneous nature of the datasets, which include a combination of structured and unstructured data formats such as text, photos, and numerical values. The presence of heterogeneity poses difficulties in both the training and the interpretation of models. Furthermore, complications arise from factors such as imbalanced class distribution, noisy data, a high number of dimensions, and changing data distribution, which further increase the complexity. Class imbalance can result in biased models, noisy data can mislead the training process, high dimensionality requires significant processing resources, and dynamic data distribution poses obstacles to the adaptation of ML models over time.

To tackle the intricacy of data, one must employ sophisticated approaches such as improved preprocessing methods, meticulous feature engineering, and the integration of flexible model architectures such as deep learning. The widespread use of ML models in many applications highlights the importance of efficiently handling data complexity to ensure the resilience and adaptability of these models in real-life situations. As artificial intelligence progresses, it is essential to conduct continuous research and development to find solutions that can reduce the effects of data complexity on the performance and dependability of ML systems.

2.0 Literature Review

The literature study begins by evaluating conventional methods of calculating insurance rates, highlighting their dependence on historical data rather than incorporating driving behavior. The characteristics emphasized that influence the risk profiles of fleet-managed vehicles include vehicle age, type, and driving behaviors. This establishes the foundation for a reassessment of traditional grading factors, with a focus on fairness and flexibility. The conversation shifts to influential literature on pay-as-you-drive (PAYD) insurance, progressing to more sophisticated approaches such as Manage How You Drive (MHYD) and Pay How You Drive (PHYD).

The investigations explore the intricate correlation between mileage, driving patterns, and the likelihood of accidents. The profound impact of usage-based insurance (UBI) is emphasized, since empirical data shows improved driving behavior and decreased accident frequencies. Although UBI offers advantages, it encounters obstacles such as privacy issues and the expenses associated with its implementation. The study anticipates that the auto insurance market would see significant changes due to the implementation of UBI, such as increased competition and structural transformations.

Traditional methods for setting insurance rates rely on frequency and severity models that utilize historical data stored in an insurance company's database to estimate the anticipated number of claims and their associated costs. In the past, information on driving behavior was not directly factored into these models because it was difficult to objectively measure driving style and intensity.

Research has shown that age and vehicle type play an important role in determining the risk profile of fleet-managed vehicles. Older vehicles are more likely to break down, require repairs, and

pose a higher risk of accidents. Similarly, vehicles with larger engines and those used for high-mileage purposes are more likely to experience accidents and incur higher repair costs. The driver experience is another key factor that impacts the risk profile of fleet-managed vehicles. Research has shown that drivers with more experience are less likely to be involved in accidents, while new or inexperienced drivers are at a higher risk of accidents.

In addition, driving patterns are an important factor in determining the risk profile of fleet-managed vehicles. Vehicles used for high-mileage purposes or those driven in urban areas are more likely to experience accidents or damage. Similarly, vehicles used for off-road purposes or in hazardous conditions are at higher risk of accidents and damage.

A decent rating variable must adhere to the following criteria:

i) It must be equitable to all.

ii) It must adhere to state regulations.

iii) It must be adaptable to changes in the policyholder's driving habits.

There is an urgent need for change because some of the conventional rating variables do not meet these requirements. First, insurance company rates must not be too high, too low, or discriminatory. In other words, the insurer charges each person a fair rate. However, many rating variables simply record a broad pattern without identifying exceptional cases within the group. For certain policyholders, the premium becomes unfair.

The work of Boucher and Turcotte (2020) and Paefgen et al. (2014), who explore risk modeling based on pay-as-you-drive (PAYD) insurance data, is among the early advances in telematics data that are more closely tied to insurance challenges. PAYD insurance is a type of usage-based insur-

ance (UBI), where the amount of the premium is determined by the number of miles driven. PAYD involves the use of telematic devices, such as GPS trackers and accelerometers, to collect data on a driver's behavior, such as their driving speed, distance traveled, and time spent on the road. This data is then used to calculate a personalized insurance premium for the driver.

Later, UBI evolved into increasingly complex strategies such as Manage How You Drive and Pay How You Drive (Arumugam and Bhargavi, 2019). PHYD determines premiums based on driving behaviors such as speeding, hard braking, and acceleration, while MHYD goes one step further by offering real-time notifications to help drivers reduce their risk of collision. The relationship between insurance rates and driving behavior that PHYD and MHYD offer may be a useful tool for reducing dangerous driving behaviors and improving road safety.

These preliminary studies aim to investigate the relationship between mileage and accident risk. According to Boucher et al. (2013), the relationship between mileage and accident risk is not linear. This might be the case because drivers who have driven more miles have likely driven on safer roads more frequently, in newer cars, and with greater driving experience. In their 2014 study, Paefgen et al. (2014) aggregated mileage according to daylight, weekday, road type, and speed intervals to examine multivariate exposures.

The Weibull regression model, sometimes referred to as the Weibull proportional hazards model, is a statistical tool utilized to analyze the duration until a specific event, such as failure or occurrence, takes place. Ayuso et al. (2016b) uses a Weibull regression for the distance traveled to the first accident at fault to create a survival model. They discover that driving at night and at high speeds shortens the time to the first collision. Using the same survival model, Ayuso et al. (2016a) discovered that the primary factor contributing to gender differences in accident risk is the frequency

with which a person uses a vehicle, that is, the fact that men drive more frequently than women. Ayuso et al. (2019) also includes information on driving habits, such as the proportion of miles driven at night, above the speed limit or in urban areas.

Research has also shown that UBI has the potential to significantly reduce the cost of motor insurance for drivers, particularly for those who drive less frequently and/or exhibit safer driving behavior. According to Zhu (2017), participants in the UBI program improve their driving performance during enrollment and obtain permanent discounts that are on average 12% lower than what they would have paid if they had not joined the program. The figure below highlights how telematics works in the insurance industry:

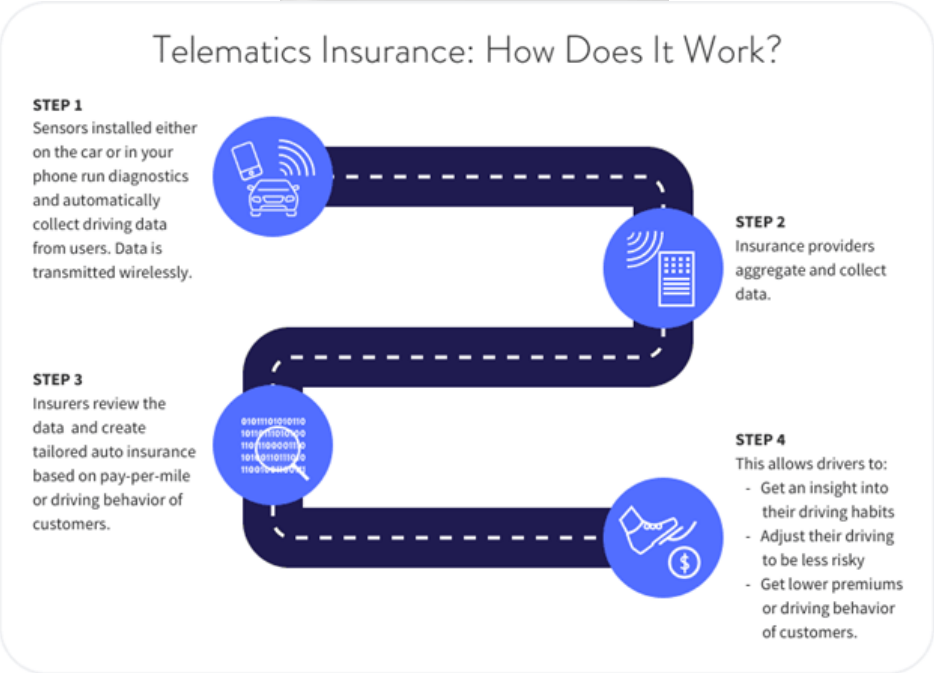


Figure 2.1: Illustration on Telematics

In addition, UBI can also induce safer driving behavior by providing drivers with feedback on their driving performance. Several studies have shown that providing drivers with feedback on their driving behavior can lead to a significant reduction in accidents and risky driving behavior.

An additional benefit for customers in terms of safer driving behavior is a reduction in the cost of fuel consumption. Acceleration, sudden braking, and sudden acceleration can lead to 40% more fuel consumption. If people drive less aggressively under the supervision of telematics devices, their fuel costs are lower. Speed is another factor that affects fuel efficiency. According to Yao (2018), the figure below shows the relationship between speed and miles per gallon. At a speed of 55 mph, the vehicle has a consumption level of 33 mpg. However, when the driver exceeds 65 mph, MPG drops sharply and fuel economy improves.

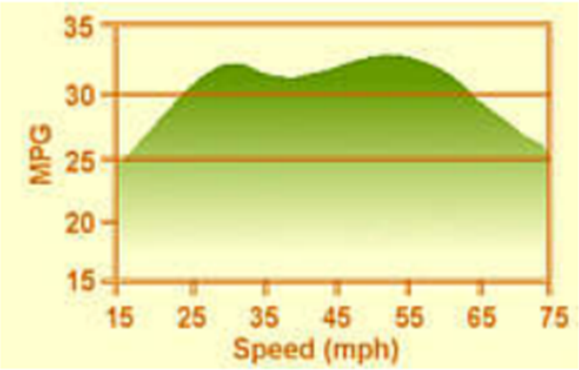


Figure 2.2: Speed vs Miles per Gallon

In addition, uninterrupted data transmission between the device and the insurance company allows insurance companies to receive an initial notice of a loss within minutes or seconds and allows them to alert crucial agencies such as police, towing firms, and other parties in the event of a loss (Palmer, 2016).

Previous research has focused mainly on PAYD characteristics, but we should also examine driving abilities and behavior in addition to driving habits. This has led to a comparison between pay-as-you-drive (PAYD) and pay-how-you-drive (PHYD) insurance plans. Such PHYD devices may be created with risky driving techniques in mind, disobeying traffic laws pertaining to speeding and using a smartphone while driving. Huang and Meng (2019) claim that the frequency prediction

model takes into account driving performance, key accidents, and travel habits. For every driver, they extract 30 telematics variables based on detailed feature engineering.

In cases where there is limited access to accident or claim data, Sun et al. (2020) utilize average accelerator pedal position and brake count as dependent variables to measure driving risk. As independent variables, they employ driving distance, speed, and rotations per minute (RPM). So et al. (2021) employ a cost-sensitive multi-class AdaBoost algorithm to estimate the frequency of accidents by analyzing the intensity of abrupt braking, acceleration, and turning, as well as the proportions of durations on various road types and during the day. It is possible to address the issue of accident class imbalance using this suggested algorithm. Denuit and Trufin (2019) have put up a credible strategy to include information related to driving experience in insurance pricing.

The primary benefit of UBI is that it allows insurers to better tailor insurance policies to individual drivers based on their actual driving behavior. Traditional insurance models rely on assumptions and statistical models to predict a driver's risk, which can result in higher premiums for low-risk drivers and lower premiums for high-risk drivers. In contrast, UBI allows insurers to reward safe drivers with lower premiums and encourage drivers to adopt safer driving habits. Several studies have shown that UBI can lead to significant reductions in accident rates and associated costs. For example, a study by Mungai and Odhiambo (2021) on UBI adoption in Nairobi's matatu or public transport sector highlighted a 30% reduction in collision frequency over a six-month telematics trial.

Despite its potential benefits, UBI faces several challenges that can hinder its widespread adoption. One of the main challenges is privacy concerns related to the collection and use of driver data. Some consumers may be uncomfortable with the idea of having their driving behavior monitored

and shared with insurance companies. To address these concerns, insurers must be transparent about the data they collect and how they use it, and offer clear opt-out options for drivers who do not wish to participate in UBI programs.

Another challenge is the high cost of implementing UBI programs, which may require significant investment in telematics technology and data engineering infrastructure. Insurers must also develop effective methods for interpreting and analyzing the vast amounts of data generated by UBI programs to accurately assess risk and set premiums. In the PSV sector, these concerns have already been tackled by companies that have properly managed systems whose fleet must be monitored for various business and regulatory reasons.

The adoption of UBI is expected to have a significant impact on the auto insurance industry. UBI is likely to lead to increased competition among insurers, as they compete to offer the most attractive premiums and rewards to safe drivers. This may result in lower premiums for drivers who adopt safer driving habits, but also better selection along with the ability to price discriminate among customers as insurers adjust to the new UBI paradigm. UBI may also lead to changes in the way insurance policies are structured and sold. Traditional policies are typically sold on an annual basis, but UBI policies may be sold on a pay-as-you-go basis, with premiums adjusted in real time based on driving behavior. This could lead to greater flexibility and affordability for drivers, particularly those who drive less frequently or for shorter distances. This study seeks to harness the benefits of the UBI model and apply it to fleet-managed vehicles in Kenya.

UBI is an innovative approach to auto insurance that offers several potential benefits, including lower premiums for safe drivers and reduced accident rates and associated costs. However, UBI also faces several challenges related to privacy concerns, high implementation costs, and the need

for an effective data engineering infrastructure. Despite these challenges, UBI is expected to have a significant impact on the auto insurance industry, leading to increased competition and changes in the way insurance policies are structured and sold.

3.0 Research Methodology

This dissertation aims to develop a user-based risk assessment framework for auto insurance in Kenya, specifically tailored for fleet managed through telematics and machine learning. The dataset comprises rating variables that influence user-based pricing collected by carlog installed in buses. The Carlog is a device in the wider telematics field that is installed in a vehicle to monitor and transmit data on its status to a remote server. The Carlog is seamlessly integrated into the vehicle's electronics to track various performance metrics, including speed, odometer readings, braking system functionality, fuel consumption, and remaining fuel levels in the tank. When connected to the vehicle's electronic display system through a CANBUS adaptor, it enables the replication of the vehicle's dashboard on a remote computer as shown in the figure below. This provides controllers with a comprehensive real-time overview of the vehicle's status (Mwithimbu, 2024). Before each trip, drivers are required to log in using a unique identifier, this is necessary because the drivers are rotated through the fleet of buses.

Figure 3.1: A snapshot of a carlog vehicle dashboard display

The system replicates the vehicle dashboard on the computer displaying the RPM, Speed, Odometer reading, Fuel status and the engine temperature. The map shows the vehicle location on a Google map. The collected data is aggregated to generate a driver score on a 100%, which considers various parameters such as operating time, geographical locations traversed, and vehicle speed.

These scores serve as comprehensive indicators of driving behavior and encompass variables such as driver identification, trip duration, distance covered, instances of over-speeding and over-RPM, idle time, acceleration, deceleration, curve acceleration, and basic vehicle metadata. The dataset represents telematics data collected from approximately 173 drivers operating across multiple

routes managed by Easy Coach Limited. Spanning the period from January 2022 to February 2024, the raw dataset comprised over 20 million rows of granular trip-level data. Due to the vast volume and high dimensionality of the collected data, it was aggregated and transformed into composite driving behavior scores to facilitate meaningful analysis, reduce computational complexity, and improve interpretability. The table below explains how the scores were determined:

Table 3.1: Variables Definition

Key Variables	**Definitions**
ID	Driver ID
Distance	GPS recorded distance aggregated to monthly intervals in Kilometres.
Duration	Duration taken to cover the distance in hours and minutes.
Over-Speeding	A score that is calculated by penalty offset when speed limit is exceeded.
Over-RPM	A score that is calculated by aggregating Revolutions per Minute recorded by a sensor at intervals.
Idle Time	A score collected by the sensor when the engine runs but the accelerator position is at 0.
Acceleration	A score that is calculated by a weighted score depending on the acceleration of the driver for the duration of a trip with 0% being poor and 100%- being excellent performance.
Deceleration	A score that is calculated by a weighted score depending on the braking of the driver for the duration of a trip with 0% being poor and 100%- being excellent performance.

3.1 Ethical Considerations

The study rigorously adheres to ethical data protection principles by implementing comprehensive anonymization protocols. All personally identifiable information (PII) including driver names, contact details, vehicle identification numbers, and geolocation coordinates were systematically excluded from the dataset prior to analysis. This pre-processing measure ensures compliance with GDPR Article 17 (Right to Erasure). The anonymization framework follows the k-anonymity model (where $k \geq 5$), guaranteeing that no individual driver can be identified through linkage attacks or quasi-identifiers. These precautions maintain the dataset's analytical utility while eliminating re-identification risks.

3.2 Data Pre-Processing

3.2.1 Data Cleaning

A crucial phase in the data analysis process is data cleaning, often known as data wrangling or preparation. To guarantee the quality and dependability of the data used for analysis, it involves locating and fixing errors, inconsistencies, and inaccuracies within the data sets. The procedure is crucial to reduce bias, improve the accuracy of insights obtained from the data, and convert raw data to a format that can be used.

If left unchecked, missing values, duplicate records, or erroneous entries in a common dataset might skew the analysis findings. Cleaning the data guarantees that it satisfies certain quality requirements, such as validity, accuracy, consistency, and completeness. In the age of big data, this procedure is particularly crucial because the amount, diversity, and speed of data can increase

errors and inconsistencies.

3.2.2 Missing Values

Addressing missing values is a common obstacle in datasets, which can undermine the effectiveness of machine learning models. Various strategies have been developed to address this challenge and maintain the reliability of machine learning algorithms trained on numerical input. The isnull() function from the pandas package is a frequently used method in python to identify missing values. This function produces a boolean mask that indicates the presence or absence of each value in the dataset. The columns that comprised 100% NaN values were removed.

3.2.3 Removing Duplicates

Removing duplicates ensures the accuracy and integrity of datasets by eliminating redundant entries. Duplicates often arise from overlapping sources, repeated data collection efforts, or errors during data entry. If not addressed, they can skew statistical calculations, misrepresent trends, and inflate both storage needs and computational costs. The process begins with defining what constitutes a duplicate, which may include identical records across all fields or specific columns, such as user IDs. Detection tools like python's pandas library or SQL queries can efficiently flag repeated entries, while validation ensures duplicates are indeed erroneous and not intentional. Once verified, duplicates can be removed using automated methods like pandas.DataFrame.drop_duplicates in python or SQL commands, ensuring only the most accurate and complete records are retained. Best practices include backing up original data, understanding the context of duplication. By removing duplicates, it created cleaner datasets that yield more reliable insights and facilitate efficient decision-making processes (Dasu and Johnson, 2003).

3.2.4 Identifying Outliers

Finding outliers is a crucial part of data cleaning since these odd or extreme results can have a big influence on the precision and dependability of analysis. Errors in data entry, inaccurate measurements, or actual anomalies in the data can all produce outliers. Their existence has the potential to skew statistical computations, including mean, variance, and correlations, producing inaccurate conclusions or predictions (Aggarwal, 2017).

Understanding the dataset and its expected range of values is the first step in the outlier detection procedure. Statistical methodologies, visualization tools, and computational approaches are common ways to find outliers. Measures like the standard deviation and the interquartile range (IQR) are often used in statistical procedures. For instance, values that fall below the first quartile or above the third quartile by more than 1.5 times the IQR are often regarded as outliers (Tukey, 1977). In this case, for distances less than 35 kilometers pointed to mechanics and inspectors and greater than 50,000 kilometers pointed to user 'Bypass' who is not a driver in the system, were removed from the dataset.

3.2.5 Oversampling using ADASYN

Machine learning algorithms face significant hurdles when trained with uneven data sets. Forecasts show misleading precision and are distorted. The lack of information related to the minority class is a fundamental factor contributing to this issue. Machine learning algorithms frequently employ the practice of classifying each test case sample into the majority class to enhance the accuracy measure. This approach is based on the premise that data sets are balanced, with equal class weights. One solution to address this issue is the utilization of sampling techniques.

As illustrated in Figure 3.2, the discrete classes were found to be imbalanced with 91.94% representing the data points in the majority class and 8.06% representing the data points in the minority class.

Figure 3.2: Class Distribution

According to Salehi and Khedmati (2024), in the case of numerous base classifiers, an imbalanced dataset tends to result in a lower overall classification performance compared to a balanced dataset. Undersampling is a straightforward approach to address the issue of uneven data. The majority class, Class 0, is undersampled. Before achieving data balance, the new training data set consisted of randomly selected data points from the majority class.

In highly imbalanced classification problems, machine learning models often become biased toward the majority class, resulting in poor generalization on the minority class. One effective strategy to address this issue is through synthetic oversampling, which aims to augment the dataset by generating artificial examples of the minority class. A prominent and adaptive technique within this category is ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning), introduced by He et al. (2008).

Unlike traditional methods such as SMOTE, which generate synthetic samples uniformly for all minority instances, ADASYN focuses its resampling efforts on *hard-to-learn* data points—those

located in regions where the minority class is surrounded by majority class instances. The central idea behind ADASYN is to adaptively generate more synthetic data for these difficult instances, effectively shifting the decision boundary in favor of the minority class while reducing bias and variance.

Mathematically, ADASYN first identifies the k-nearest neighbors of each minority class sample and calculates a density distribution that reflects how many majority class samples surround it. The number of synthetic samples to be generated for each instance is proportional to this local learning difficulty. Synthetic instances are then created by interpolating between the minority instance and its selected minority neighbors, with a random weighting factor.

This mechanism allows ADASYN to achieve two goals simultaneously:

i. Enhance the learning focus on minority samples that are harder to classify.

ii. Avoid over-generalization by reducing the creation of redundant or overlapping minority instances.

By generating targeted and context-aware synthetic examples, ADASYN enables classifiers to learn more precise decision boundaries, particularly in severely imbalanced scenarios where the minority class constitutes a very small fraction (less than 10%) of the data. This makes it especially suitable for tasks in domains such as fraud detection, rare disease diagnosis, and anomaly detection, where correct minority classification is crucial.

3.3 Exploratory Data Analytics

Exploratory Data Analysis (EDA) is a crucial stage in the data analysis process that facilitates the

detection of patterns, anomalies, and correlations within data. This procedure enables analysts to formulate hypotheses and enhance modeling decisions. EDA is situated inside the comprehensive data life cycle, occurring between data collection and preprocessing at one end, and modeling, validation, and deployment at the other. By meticulously examining the data at this preliminary phase, exploratory data analysis enhances the quality of ensuing analytical and modeling procedures.

The EDA is essential for comprehending the distinct attributes of data, detecting errors and inconsistencies, and uncovering correlations among variables. It also aids in validating assumptions and guiding judgments regarding the most appropriate models for the analysis. Tukey (1977), played a crucial role in the advancement of Exploratory Data Analysis (EDA), highlighting visual exploration as a fundamental approach to uncovering insights in data. His work underlies numerous contemporary procedures that employ graphical methods to enhance the accessibility and interpretability of complex data.

The EDA procedure commences with a meticulous analysis of the dataset's population and the data dictionaries to comprehend the definitions and classifications of variables. Subsequently, summary statistics and data visualizations are employed to derive insights into the structure, central tendencies, and variability of the data set. This analysis entails analyzing correlations among variables and detecting and evaluating missing data or outliers that may impact the integrity of the analysis. This thorough methodology allows analysts to discern essential elements that must be taken into account in future modeling stages.

The analysis utilizes an array of tools and methodologies, such as summary statistics, data visualization, and dimensionality reduction, to assist analysts in acquiring preliminary insights into their data. These strategies elucidate data trends and key tendencies while also enhancing compre-

hension of diversity within the dataset. Utilizing these tools, the data was prepared for modeling successfully, so ensuring a more robust and precise study. Descriptive statistics were utilized to succinctly summarize the key elements of driving behavior data. Essential metrics, including the mean speeding score, the median frequency of idle time incidents, and the variability in scores, are computed. These indicators provide significant insights on current driving behaviors. Visual instruments like histograms and box plots are widely employed to depict the distributions of various variables, therefore aiding in the detection of outliers and abnormalities that may distort the analysis. The descriptive statistics indicated that certain data points were omitted due to mechanics and inspectors traversing distances considered minor and non-risky.

3.3.1 Feature Engineering

Feature engineering plays a critical role in improving the predictive capabilities of ML models by transforming raw data into informative indicators. The features in the dataset are presented as scores that are developed using a weighted score or penalty offset. In the context of developing an over-speeding score, the process involves quantifying driving behavior through a structured penalty offset system based on speed deviations from legal limits on different types of roads. For instance, speeds within 5 mph of the limit on highways and streets incur minor offsets, while larger deviations result in more significant penalties. This systematic approach allows for the nuanced assessment of driving habits, distinguishing between minor and more severe instances of over-speeding by applying differentiated penalty scores for varying degrees of speed limit breaches as explained in Table 3.2 below:

Table 3.2: Over-Speeding Feature Engineering

Speed	**Highway**	**Streets**	**Penalty Score**
Speed limit +/- 5	+0.015	+0.025	-
Speed limit +/- 15	+0.05	+0.10	-
Driver A	10/4	4/2	0.65
Driver B	2/1	8/0	0.28
Driver C	5/0	5/0	0.2

Calculating the speeding override score for drivers, as illustrated in Drivers A, B, and C, utilizes a quantitative method that aggregates penalty offsets from multiple speeding events. This method evaluates each driver's adherence to speed limits by accounting for the frequency and severity of over-speeding incidents. For example, Driver A's score is derived from a combination of minor and major speeding events on highways and streets, leading to a cumulative penalty of 0.65. This granular analysis not only captures the frequency of over-speeding occurrences, but also emphasizes the importance of contextual factors such as road type in assessing driving behavior (Warren and Greenlee, 2006). Through this feature engineering technique, the overspeeding score emerges as a dynamic and insightful component of the overall driver evaluation system, reflecting both the intensity and regularity of over-speeding incidents.

At the end of the month, the scores are subsequently combined to calculate a cumulative value, which is subsequently used to categorize drivers into different segments for each variable. The scores are then summed up and the result is designated as the target variable.

$$\text{Total Driver Score} = \left(\frac{\sum_{i=1}^{n} z_i}{n} \right) \times 100 \quad (1)$$

Where:

z_i = the total weight obtained by all attributes,

n = Number of attributes observed.

3.3.2 Univariate, Bivariate, and Multivariate Analysis

The investigation extends to the identification of patterns within driving behavior through univariate, bivariate, and multivariate analyses. The univariate analysis focuses on the frequency of individual variables, uncovering correlations between acceleration and deceleration scores and the company's peak activity periods. An interesting observation is the decline in scores during high demand periods, coinciding with the academic calendar in Kenya especially in August, hinting at potential opportunities for the company to optimize resource allocation during these times.

Bivariate analysis reveals the distribution of idle time across the dataset, indicating a generally positive idle-time behavior among drivers. This suggests efficient driving practices, with minimal unnecessary idling. A scatter plot examining the relationship between idle time and acceleration uncovers a weak positive correlation, suggesting that drivers in congested traffic conditions might exhibit quicker acceleration patterns.

Multivariate analysis through a correlation matrix brings to light several significant relationships among variables. In particular, a strong positive correlation exists between the distance traveled and the duration of the journey, as well as between instances of overspeeding and the overall

score, indicating the impact of such behaviors on driving performance. In addition, a moderate positive correlation between acceleration and deceleration suggests that these behaviors often occur simultaneously, underscoring the complex dynamics of driving patterns.

This comprehensive exploratory data analysis provides a foundation for further in-depth study, highlighting the intricate interplay of various driving behaviors and their collective influence on driving performance and safety. Further exploration of these relationships is essential to develop more effective driving practices and insurance models.

3.4 Machine Learning Modeling

This section will look at the behavioral analysis of drivers and classify them according to their driving habits. The analysis employs three machine learning techniques and statistical indicators are used to compare how well each performs. To demonstrate how massive data-driven machine learning algorithms are applied in the sector, supervised machine learning algorithms will be used. In this work, machine learning algorithms that include logistic regression, random forest, and gradient-boosting trees were chosen.

Logistic regression is often used in the insurance sector for binary categorization because of its straightforward interpretability. However, its capacity to capture intricate non-linear interactions is limited. To overcome this constraint, the study incorporated advanced techniques such as random forests and gradient-boosting trees to reveal the intricate connections present in the data. The ensemble classifier known as random forests addresses the constraints of individual decision trees by combining the results of numerous trees. Random forests typically surpass decision trees alone by addressing overfitting. Gaussian boosting trees, which are a type of ensemble classifier, generate

a prediction model by combining many weak decision trees. Random forests are typically outperformed by this strategy. The subsequent section provides an elaborate and thorough elucidation of each of the aforementioned methodologies, exploring their fundamental concepts and complexities. In addition, a concise explanation of the execution of these techniques is provided to give a pragmatic understanding of their use within the context of this research.

3.5 Logistic Regression

Logistic regression is a widely used statistical method for analyzing binary classification problems, where the outcome variable takes one of two possible values, typically coded as 0 or 1. Unlike linear regression, which assumes a continuous response, logistic regression models the probability that the response variable Y belongs to one of the two classes. According to James et al. (2013), the logistic regression model is particularly suitable when the goal is to predict a binary response and estimate the probabilities of class membership.

In logistic regression, the relationship between the predictor variables $X = (X_1, X_2, \ldots, X_p)$ and the binary outcome Y is modeled using the logit function, which transforms the response probability into a linear function of the predictors. Specifically, the log-odds (logit) of the probability $P(Y = 1)$ are expressed as:

$$\text{logit}(P(Y=1)) = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{2}$$

where β_0 is the intercept, and $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients associated with the predictor variables. This linear representation of the log-odds makes logistic regression interpretable and com-

putationally efficient.

To transform the log-odds back into a probability, the logistic function (sigmoid function) is applied:

$$P(Y=1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)} \tag{3}$$

This function ensures that the predicted probabilities fall within the interval $[0,1]$. The coefficients β_j are estimated using Maximum Likelihood Estimation (MLE), a method that finds the values of the parameters that maximize the likelihood of observing the given data.

Logistic regression is advantageous in its interpretability, as the coefficients can be directly linked to changes in the log-odds. For example, a one-unit increase in a predictor X_j corresponds to an increase of β_j in the log-odds, holding all other predictors constant. This property makes logistic regression particularly useful in applications where understanding the relationship between predictors and response is as important as prediction accuracy.

James et al. (2013) highlight that logistic regression is effective in many real-world scenarios, including medical diagnosis, credit scoring, and risk assessment. However, logistic regression assumes a linear relationship between the log-odds of the outcome and the predictors. If this assumption is violated, transformations of the predictors or inclusion of interaction terms may be necessary to improve the model performance. Furthermore, logistic regression can suffer from multicollinearity when the predictor variables are highly correlated, which can affect the stability of coefficient estimates.

In this research, logistic regression will be applied to model the binary response variable and assess the effects of multiple predictors. The model will be evaluated using metrics such as accuracy,

precision, recall, and the area under the ROC curve (AUC) to ensure its robustness and predictive power. Furthermore, the statistical significance of the coefficients will be examined to identify the most influential predictors in the dataset.

3.6 Random Forest

Random Forest is a powerful ensemble learning method introduced by Breiman (2001) combines multiple decision trees to improve predictive accuracy and control overfitting. It operates by creating a "forest" of decision trees, each trained on a random subset of the training data and a random subset of the predictor variables. The final prediction is obtained through aggregation, such as majority voting for classification or averaging for regression.

The algorithm begins by generating multiple bootstrapped samples from the original dataset, where each sample is created by randomly selecting the observations with replacement. For each bootstrapped sample, a decision tree is constructed considering a random subset of predictors at each node to determine the best split. This randomization ensures that the trees in the forest are diverse, which reduces overfitting. The trees are grown to their maximum depth without pruning, further highlighting their independence. Once the individual trees are built, the predictions are aggregated. For classification tasks, the model uses majority voting to determine the predicted class:

$$\hat{y} = \text{majority\_vote}\left(T_1(x), T_2(x), \ldots, T_m(x)\right) \tag{4}$$

where $T_1, T_2, \ldots, T_m$ are the m decision trees, and x represents the input feature vector. For regres-

sion tasks, the prediction is calculated as the average of the individual tree outputs:

$$\hat{y} = \frac{1}{m} \sum_{j=1}^{m} T_j(x) \tag{5}$$

The random forest is particularly robust against overfitting due to the combination of multiple decision trees and the randomization introduced during the training process. The ensemble approach reduces variance, and the random selection of features ensures that the model does not rely excessively on any single predictor. In addition, the method provides feature importance measures that help identify the most influential predictors in the data set. Breiman (2001) also highlights the computational efficiency of random forest, as its construction is inherently parallel, making it suitable for large datasets and high-dimensional feature spaces. Moreover, the method is less sensitive to noisy data compared to individual decision trees.

Despite its advantages, random forest has some limitations. It can become computationally expensive when a large number of trees are used, especially with very large datasets. Furthermore, interpretability is reduced because the aggregated prediction masks the decision-making processes of individual trees. For regression tasks, random forest can struggle with extrapolation as it relies on splitting rules learned from the training data.

In this research, random forest will be employed for the classification of driving behavior. The number of trees (m) and the number of predictors considered for splitting (k) will be optimized using grid search to ensure the best performance. Model evaluation metrics will include accuracy, precision, recall, and the area under the ROC curve (AUC) for classification tasks, or mean squared error (MSE) for regression tasks, ensuring a comprehensive assessment of the model's predictive capabilities.

3.7 Extreme Gradient Boosting (XGBoost)

XGBoost, or eXtreme Gradient Boosting, is an advanced implementation of the gradient boosting framework introduced by Chen and Guestrin (2016). It is designed for high performance, scalability, and efficiency in structured data prediction tasks. By building models sequentially, XGBoost corrects the errors of previous models, making it particularly effective for tasks requiring high predictive accuracy. Its robustness, speed, and flexibility have established it as a popular choice in machine learning competitions and real-world applications.

XGBoost minimizes a regularized objective function that combines two components: a loss function, which measures the predictive performance of the model, and a regularization term, which penalizes the complexity of the model to improve generalization. This objective function can be expressed as:

$$\mathscr{L}(\Theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

where $l\left(y_i, \hat{y}_i\right)$ is the loss function (e.g., squared error for regression or log loss for classification), $\Omega(f_k)$ is the regularization term for the k-th tree, f_k represents a tree in the ensemble, and Θ encompasses all model parameters. The regularization term $\Omega(f)$ is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2, \tag{7}$$

where T is the number of leaves in the tree, $\|w\|^2$ represents the L2 norm of the leaf weights, and γ and λ are hyperparameters that control the regularization strength. This term discourages overly complex trees, ensuring better generalization on unseen data.

At each iteration, XGBoost adds a new tree to the ensemble to minimize the objective function. This optimization uses a second-order Taylor expansion, which incorporates both the gradient (first derivative) and the Hessian (second derivative) of the loss function. The approximate objective for the t-th iteration is:

$$\mathscr{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i h_i f(x_i) + \frac{1}{2} h_i f(x_i)^2 \right] + \Omega(f) \tag{8}$$

where $g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$ is the gradient, $h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)2}}$ is the Hessian and $f(x_i)$ represents the prediction of the new tree, for instance i. Inclusion of second-order derivatives allows for more efficient and accurate optimization, enabling faster convergence.

XGBoost's efficiency is further enhanced through several optimizations, including sparse matrix operations, parallelized tree construction, and histogram-based split finding. These features make it suitable for large datasets and high-dimensional problems. Additionally, XGBoost can handle missing data effectively by learning the optimal splits for missing values during training. Another advantage is the ability to compute feature importance scores, which aid in understanding the contribution of individual predictors to the model's performance.

Despite its strengths, XGBoost has certain limitations. It can be computationally expensive for very large datasets, particularly when using a large number of trees or complex parameter tuning. Furthermore, like other ensemble tree methods, XGBoost can lack interpretability, as the ensemble structure obscures the decision-making process of individual trees. It may also struggle with data sets that contain excessive noise or non-informative features.

In this research, XGBoost will be applied to the classification of driver behavior. Key hyperparameters, such as the learning rate, the maximum depth of trees, and the number of estimators, will be

optimized using grid search. The performance of the model will be evaluated using metrics such as accuracy, precision, recall and F1 score for classification tasks, or mean squared error (MSE) and R-squared for regression tasks. Feature importance scores will also be analyzed to identify key predictors influencing the outcome.

3.8 Machine Learning Performance Metrics

3.8.1 Confusion Matrix

True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) statistics are used to build the confusion matrix. There are four potential results when a classifier and an instance are provided. A positive instance is considered a true positive if it is classed as positive, and a false negative if it is classified as negative. A negative instance is considered a true negative if it is classed as negative and a false positive if it is labeled as positive. A two-by-two confusion matrix, also known as a contingency table, can be created to describe the dispositions of a collection of examples given a classifier and a test set. This matrix serves as the foundation for numerous prevalent measures.

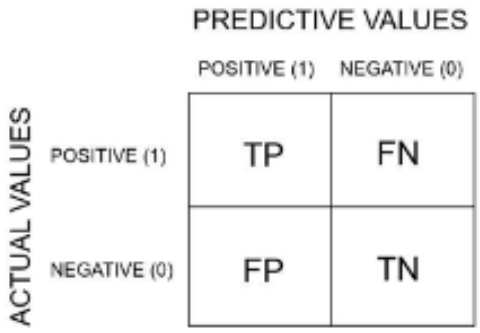


Figure 3.3: Confusion Matrix

3.8.2 Precision

The accuracy of a classification is determined by dividing the total number of valid predictions made by a model by the total number of data sets utilized in the classification.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{9}$$

3.8.3 Recall

The developed model identifies all pertinent examples or instances in the case of recollection. Recall is calculated using the equation below.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{10}$$

3.8.4 F-measure

Combining the two approaches above yields the F-measure, which is determined using precision and recall as harmonic means. Also known as F-score and F1-measure. The F-score equation is displayed as follows:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{11}$$

3.8.5 ROC Curve

According to Fawcett (2006), Rectified Receiver Operating Characteristic (ROC) graphs are graphical representations that depict the relationship between tp rate on the Y axis and fp rate on the X axis. A Receiver Operating Characteristic (ROC) graph illustrates the ratio of benefits (true positives) to costs (false positives).

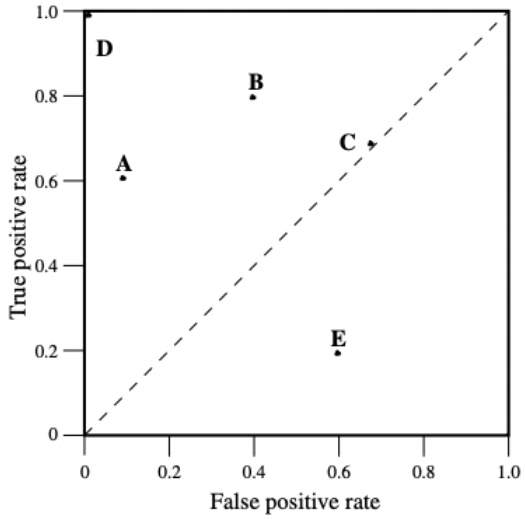


Figure 3.4: The ROC curve

A discrete classifier is a classification algorithm that is capable of generating a single class label. Each classifier produces a set of values (fp rate, tp rate) that correspond to a specific position in the receiver operating characteristic (ROC) space. Each classifier shown in Figure 3.4 is a unique classifier. Putting emphasis on some essential components within the ROC area is of utmost importance. The point (0, 0) on the lower left indicates the strategy of abstaining from producing positive categorizations. The classifier in question demonstrates a notable lack of false positive errors while simultaneously failing to generate any true positives. The upper right coordinate (1, 1) represents the counteractive method, which is defined by the unconditional assignment of

positive categorizations.

The coordinates (0, 1) indicate a state of perfect categorization. The flawless performance of "D" has been demonstrated. Informally, a point within the receiver operating characteristic (ROC) space is deemed preferable to another position if it is located to the northwest, exhibiting a greater peak rate, a lower false positive rate, or both, compared to the initial point. Classifiers located on the left side of a receiver operating characteristic (ROC) graph, near the X axis, are expected to offer clarity.

4.0 Discussion of Results

The purpose of this study is to identify a means of assessing risk accorded by insurance companies and building a driver risk classification model. The variables examined take into account driving behaviour data which includes over-speeding, harsh acceleration and sudden braking. These variables were obtained as scores that were calculated using a penalty compensation system that took into account various rules. For example, for speeding, the rules revolved around a speed limit that is set depending on the road that Easy Coach uses as its routes, the positions of the accelerator and brakes for the duration of the trips and the idle time of the engine. These scores are then evaluated using exploratory data analysis and a risk assessment model is built.

4.1 Data Pre-Processing

4.1.1 Missing Values

Columns with missing values and with more than 90% null values were removed using the .isnull() function and columns with missing values with mainly 0 such as curve acceleration and legal exceed limit were excluded from the study.

Before modeling, it is frequently required to remove zeros from a dataset for a number of reasons. To begin with, the presence of zero values can distort the distribution of the data, particularly if they are widespread or influential, resulting in imprecise representations of the data resulting in skewness. Moreover, the presence of zero values might lead to errors or undefined outcomes in mathematical computations, thus presenting difficulties in maintaining the stability of the model. The removal of zeros enables the model to focus on learning about patterns and correlations from

null values, leading to improved performance and more precise predictions. Furthermore, the elimination of zeros improves the comprehensibility of the model by diminishing extraneous information and uncertainty in the data, hence facilitating the understanding and interpretation of the outcomes. Moreover, the resolution of data sparsity concerns arising from the presence of zero values guarantees that the model acquires knowledge from significant patterns, hence enhancing the caliber and dependability of its predictions.

4.1.2 Identifying Outliers

Outliers were identified and addressed using the interquartile range (IQR) method, targeting extreme values such as distances below 35 kilometers, linked to mechanics and inspectors, and those greater than 50,000 kilometers, associated with the bypass user, which would represent driving tests. Removing these anomalies, which accounted for 5.3% of the dataset, reduced skewness and improved the reliability of statistical measures such as mean and correlation. After cleaning, clearer patterns emerged, including a significant positive relationship between idle time and travel distances, previously obscured by outliers. This process highlighted the critical role of data cleaning in ensuring accurate and robust analysis.

4.1.3 Class imbalance

Machine learning models often struggle with imbalanced datasets, where predictions exhibit misleading accuracy due to insufficient representation of the minority class. When trained on such data, algorithms tend to favor the majority class, compromising their ability to generalize. To mitigate this issue, sampling techniques are employed.

In this study, the initial class distribution was highly skewed, with 91.94% of data points be-

longing to the majority class and only 8.06% to the minority class (Figure 3.2). SMOTE was considered but did not particularly result to change in the performance of the models. For more complex imbalances, ADASYN (Adaptive Synthetic Sampling) offered a refined approach. Unlike SMOTE, which uniformly generates synthetic samples, ADASYN prioritizes "hard-to-learn" minority instances—those surrounded by majority-class samples. It adaptively adjusts synthetic sample generation based on local density, improving classifier focus on critical regions. The results of ADASYN are visualized in the figure below, showcasing its ability to rebalance the dataset while minimizing redundancy.

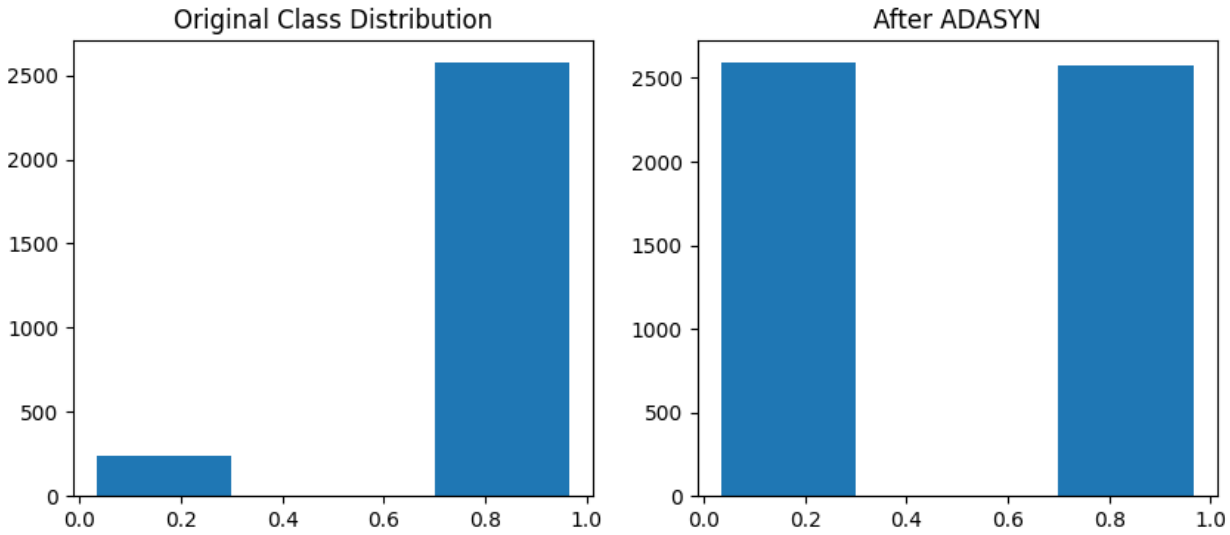


Figure 4.1: Class Distribution Before and After ADASYN

By leveraging these techniques, models achieve fairer decision boundaries, enhancing performance in critical applications like fraud detection or medical diagnosis, where minority-class accuracy is paramount.

4.2 Exploratory Data Analysis

This section looks at any patterns in driving behavior derived during the study. Univariate, bivari-

ate, and multivariate analyses are performed, and the corresponding results are presented in the subsequent sections.

4.2.1 Univariate Analysis

The analysis of acceleration and deceleration scores over time provides valuable insights into the driving performance trends within Easy Coach Ltd, particularly in relation to peak activity periods and external factors such as the academic calendar in Kenya. The data reveals a gradual decline in both acceleration and deceleration scores from April 2022 to January 2024, with average acceleration scores decreasing from approximately 99.5 to 98.5 and deceleration scores showing similar downward trends(see Figure 4.2 and Figure 4.3). This decline is particularly pronounced during peak operational periods, such as July 2022 and January 2023, which coincide with school holidays and heightened trip demand. The distribution of scores further highlights this variability, with lower acceleration and deceleration scores frequently observed during these high-demand periods. This suggests that drivers may experience increased pressure during peak times, leading to suboptimal driving behaviors such as harder braking and slower acceleration.

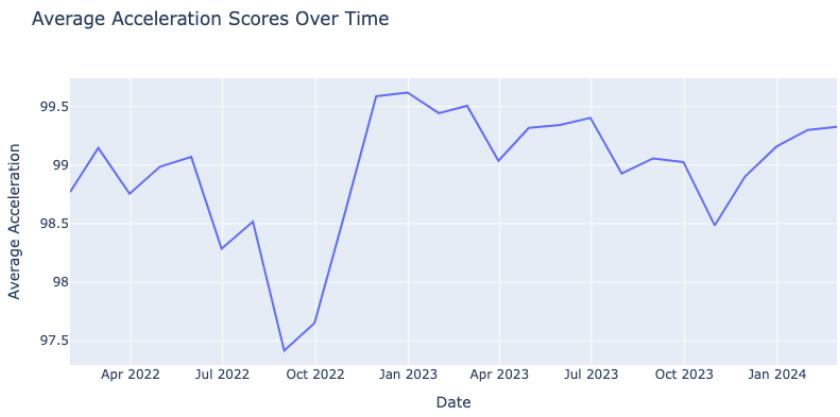


Figure 4.2: Average Acceleration over Time

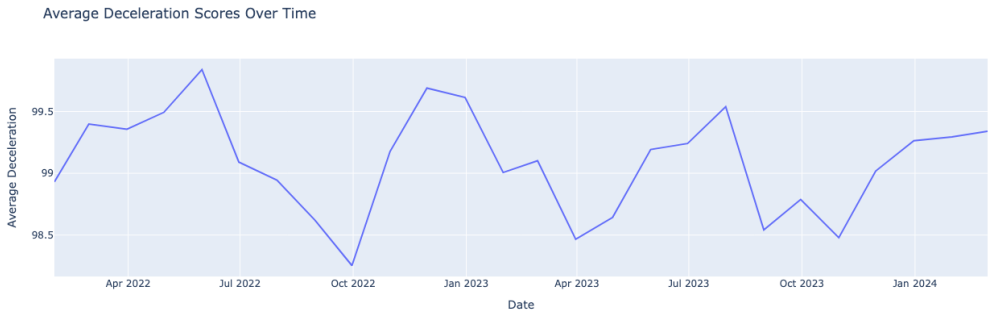


Figure 4.3: Average Deceleration over Time

The correlation between driving performance and peak activity periods underscores the impact of external factors on the Company's operations. For instance, during school holidays, the Company faces a surge in passenger demand, resulting in longer working hours for drivers and increased vehicle usage. This operational strain likely contributes to driver fatigue, which is reflected in the lower acceleration and deceleration scores. To address these challenges, the Company could implement targeted interventions during peak periods, such as driver training programs to reinforce safe driving practices, strategic resource allocation to manage increased demand, and fuel optimization strategies to reduce costs and improve efficiency. Additionally, real-time monitoring systems could be introduced to track driving metrics and enable proactive interventions when performance declines.

The graph below depicts the average speeding scores over time reveals a gradual decline in performance, with scores decreasing from approximately 95 in April 2022 to around 85 by January 2024. This downward trend indicates a deterioration in speeding-related driving behavior over the observed period. Furthermore, the graph exhibits notable seasonal variations, with significant declines in speeding scores during specific intervals. For instance, a sharp drop is observed around July 2022, coinciding with school holidays and heightened trip demand, while another de-

cline occurs in January 2023, which similarly aligns with peak activity periods. These fluctuations suggest that speeding behavior tends to worsen during high-demand periods, likely attributable to increased operational pressure and driver fatigue. Such patterns underscore the need for targeted interventions to address the root causes of speeding and improve overall driving performance.

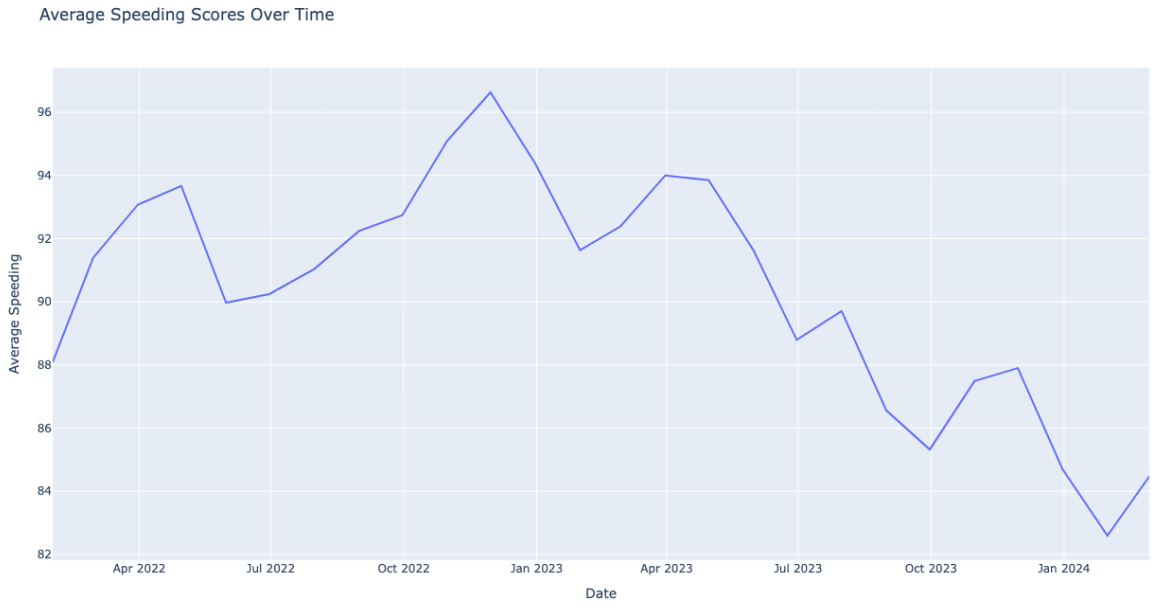


Figure 4.4: Average Speeding over Time

In conclusion, this analysis underscores the critical interplay between driving performance, peak activity periods, and external factors such as the academic calendar. The observed trends in acceleration, deceleration and speeding scores reveal significant challenges during high demand periods, likely exacerbated by driver fatigue and increased operational strain. For instance, the graph illustrates a gradual decline in average speeding scores from April 2022 to January 2024, with scores dropping from approximately 95 to 85. This decline is particularly pronounced during peak periods such as July 2022 and January 2023, aligning with the trends observed in acceleration and deceleration scores. These findings highlight the need for targeted interventions to mitigate performance

declines and optimize operational efficiency.

The adoption of a Pay-As-You-Drive (PAYD) insurance solution emerges as a strategic response to these challenges. By aligning insurance premiums with real-time driving behavior, the proposed PAYD system incentivizes safer driving practices, thus addressing the root causes of declining performance scores. This approach not only reduces insurance costs, but also fosters a culture of accountability and safety among drivers. Furthermore, the integration of real-time monitoring systems, a core component of PAYD, provides the Company with actionable insights to improve resource allocation and operational planning, particularly during peak activity periods.

4.2.2 Bivariate Analysis

The graph below depicts average idle time scores over time reveals performance fluctuations, with no consistent upward or downward trend, suggesting that idle time management has remained relatively stable over the observed period, despite periodic variations. The long-term stability of idle time scores indicates that the Company has maintained a consistent baseline performance in managing idle time, even as seasonal fluctuations occur during peak activity periods. Although idle time is not a direct factor in Pay-As-You-Drive (PAYD) insurance, reducing unnecessary idle time can contribute to lower fuel consumption and operational costs, indirectly aligning with the goals of a PAYD system. Furthermore, improved idle time management can enhance driver productivity and reduce vehicle wear and tear.

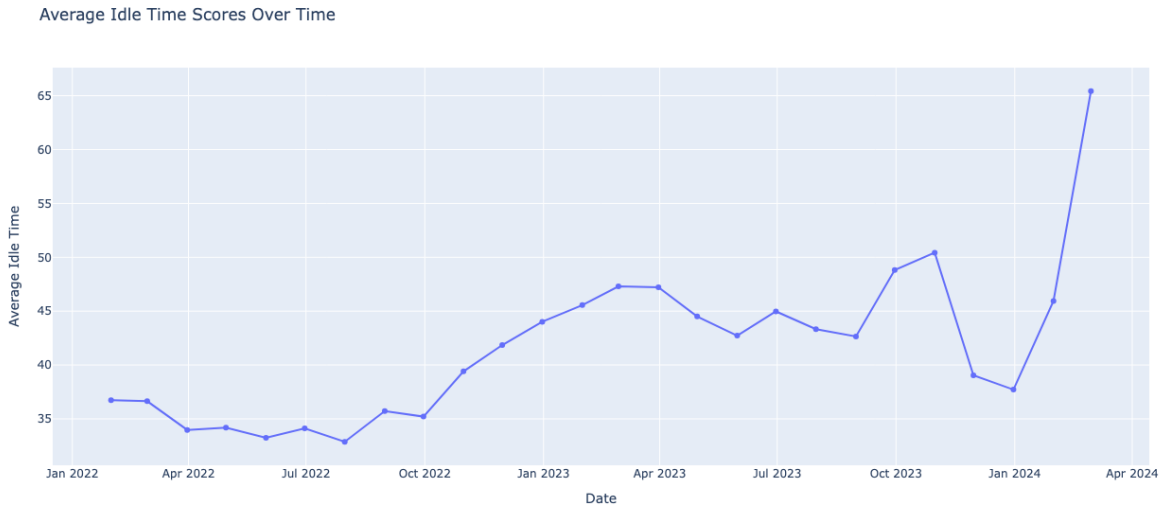


Figure 4.5: Average Idle Time

The scatter plot shows the relationship between idle time and acceleration for different drivers. The plot shows that there is a weak positive correlation between idle time and acceleration. This means that drivers who spend more time idling tend to accelerate more quickly. It could be that drivers who spend more time in traffic are more likely to idle and accelerate quickly.

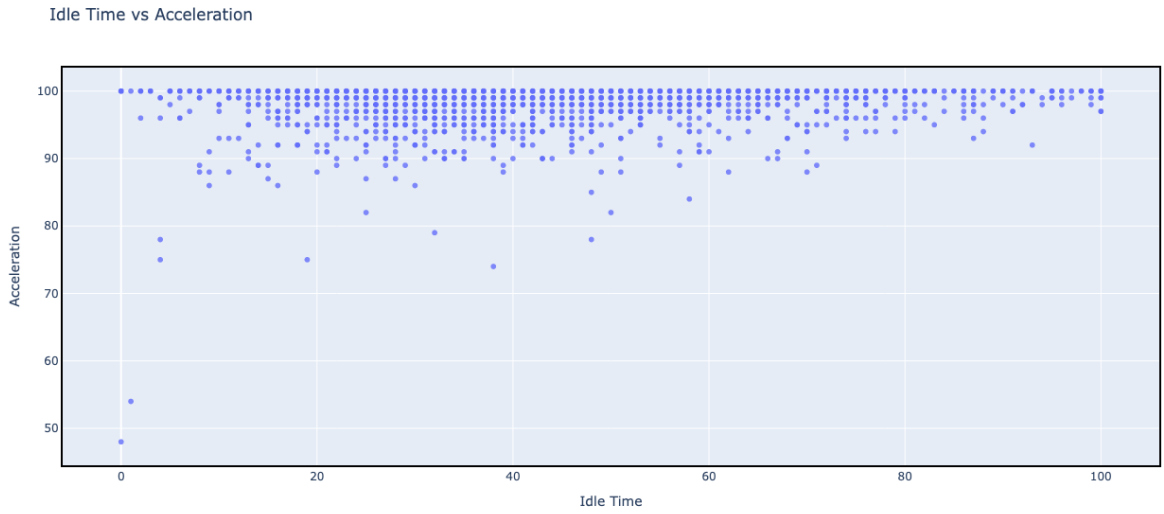


Figure 4.6: Idle Time Vs Acceleration

A scatter plot reveals a positive correlation between acceleration (Acc.) and deceleration (Dec.) scores, indicating that as acceleration scores increase, deceleration scores also tend to rise. Since both attributes contribute to controlled and responsive driving, this trend implies that drivers with strong acceleration skills are also proficient in deceleration, which is generally regarded as a positive driving behavior. This is represented as shown:

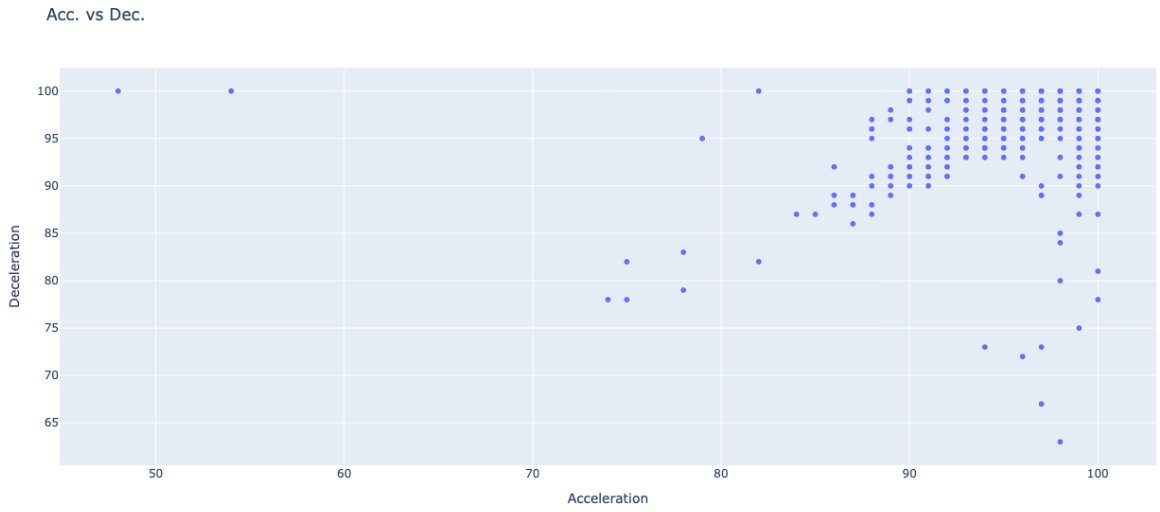


Figure 4.7: Acceleration Vs Deceleration

4.2.3 Multivariate Analysis

The correlation matrix reveals several notable relationships between variables. First, a strong positive correlation (0.906) is observed between the distance traveled and the duration of the journey, which is intuitive, as longer distances typically entail longer duration. Secondly, instances of over speeding exhibit a significant positive correlation (0.757) with the total score, indicating that occurrences of over speeding contribute to a lower overall score. Furthermore, idle time demonstrates a strong positive correlation (0.691) with the total score, suggesting that longer idle times contribute to a lower overall score. Moreover, there is a moderate positive correlation (0.174) between acceleration and deceleration, indicating that these driving behaviors tend to co-occur during driving instances. These insights shed light on the complex interaction between different driving variables and their collective impact on driving behavior and performance.

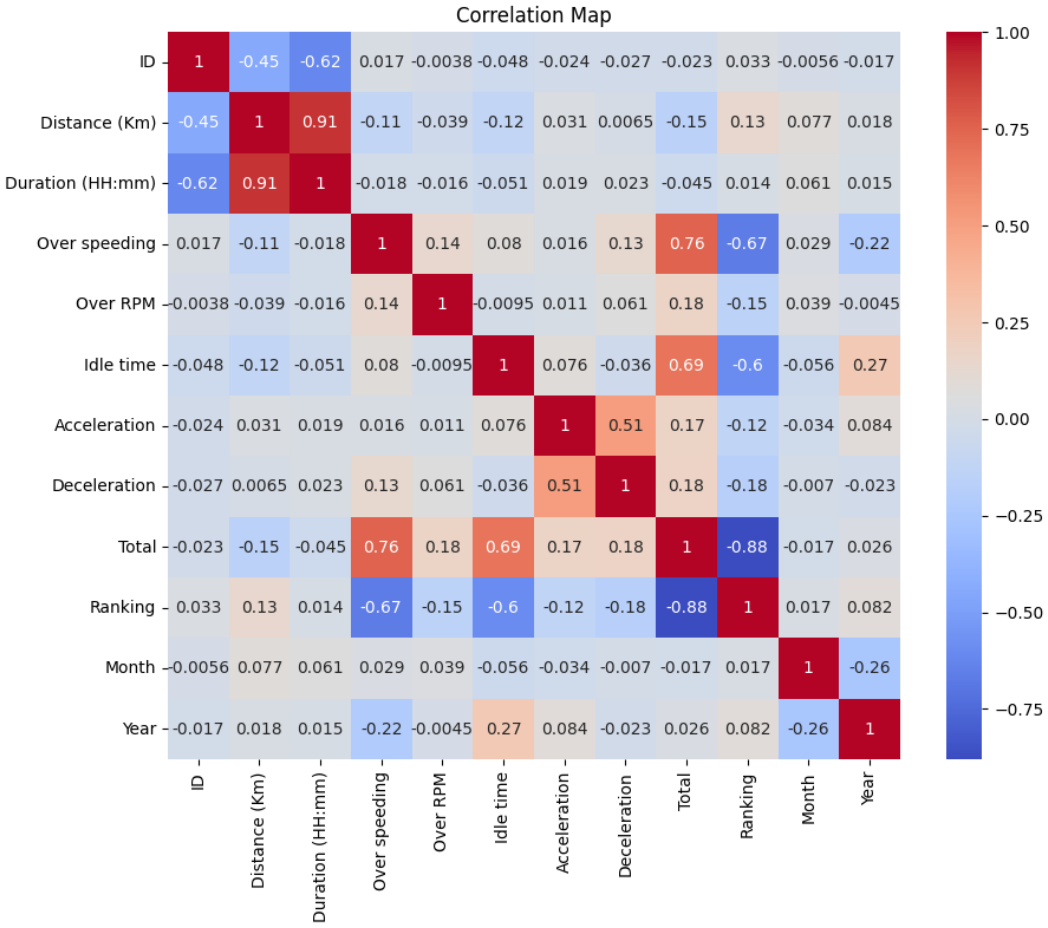


Figure 4.8: Correlation Matrix

4.3 Modeling and Performance Evaluation

XGBoost was selected as the best performing model. This is in comparison to the base model used. Since there was a class imbalance, ADASYN was applied to the dataset which improved the performance of the binary classification of the target variable. The experimental results below demonstrate that the proposed classification framework achieves exceptional predictive performance, attaining an overall accuracy of 99% on the test set.

Table 4.1: Classification Report

Class	Precision	Recall	F1-Score	Support
0	0.93	0.97	0.95	102
1	1.00	0.99	1.00	1105
Macro Avg	0.96	0.98	0.97	1207
Weighted Avg	0.99	0.99	0.99	1207

This classification rate suggests strong alignment between the model's predictions and ground truth observations of driving behaviors. Detailed analysis of class-specific metrics reveals particularly robust performance across both minority and majority classes. For the risky driving class (Class 0), the model achieves a precision of 0.93 and recall of 0.97, yielding an F_1-score of 0.95, indicating effective minimization of both false positives and false negatives. The safe driving class (Class 1) shows even stronger performance with perfect precision (1.00) and near-perfect recall (0.99), resulting in an optimal F_1-score of 1.00. These metrics collectively demonstrate the model's capacity to handle the inherent class imbalance while maintaining high discriminative power. The outstanding performance is further confirmed by the Receiver Operating Characteristic (ROC) analysis, which yields an Area Under the Curve (AUC) of 0.98, suggesting excellent separation between the two behavioral classes. This high AUC value, when combined with the strong precision-recall metrics, provides compelling evidence for the model's effectiveness as a reliable classifier for driving behavior analysis.

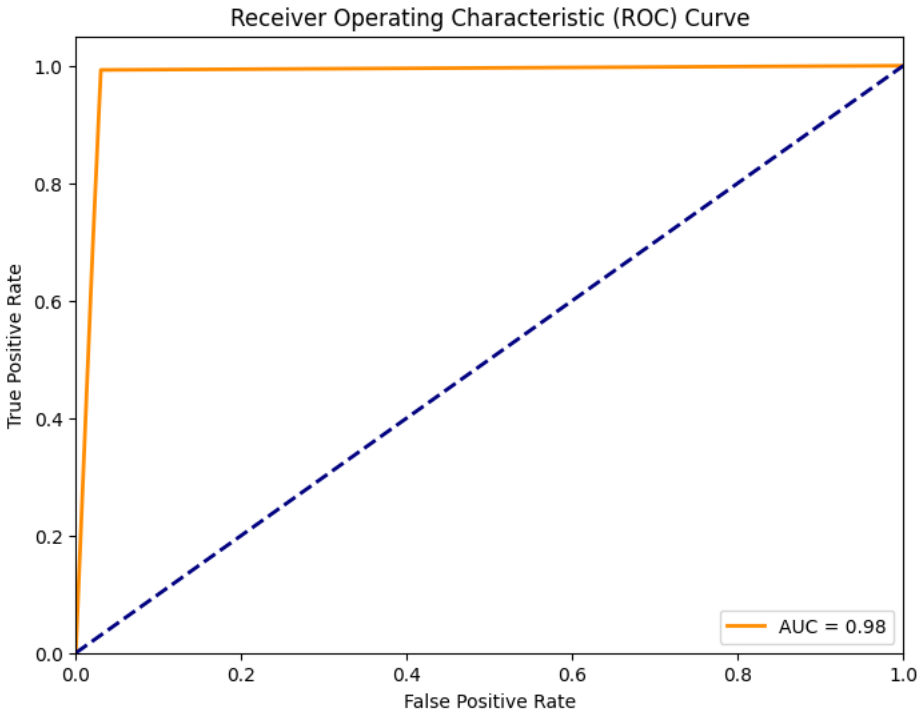


Figure 4.9: AUC Curve

The effectiveness of the model has been demonstrated in classifying driving behavior according to risk levels and discerning safe driving practices. Consequently, insurance providers have the ability to motivate low-risk drivers through the provision of premium discounts, thus fostering a culture of responsible driving. In contrast, insurers should consider charging higher premiums to high-risk drivers and offering constructive feedback to facilitate driving improvement. In summation, the classification model shows commendable performance in various evaluative metrics. These findings provide valuable information for the assessment of driving risk and facilitate informed decision making within the scope of usage-based insurance frameworks.

4.4 Deployment

The machine learning model, serialized in .pkl (pickle) format, is deployed within a cloud-based architecture on Amazon Web Services (AWS), providing secure, scalable, and monitorable inference capabilities. The deployment pipeline is visualized in Figure 4.10 below:

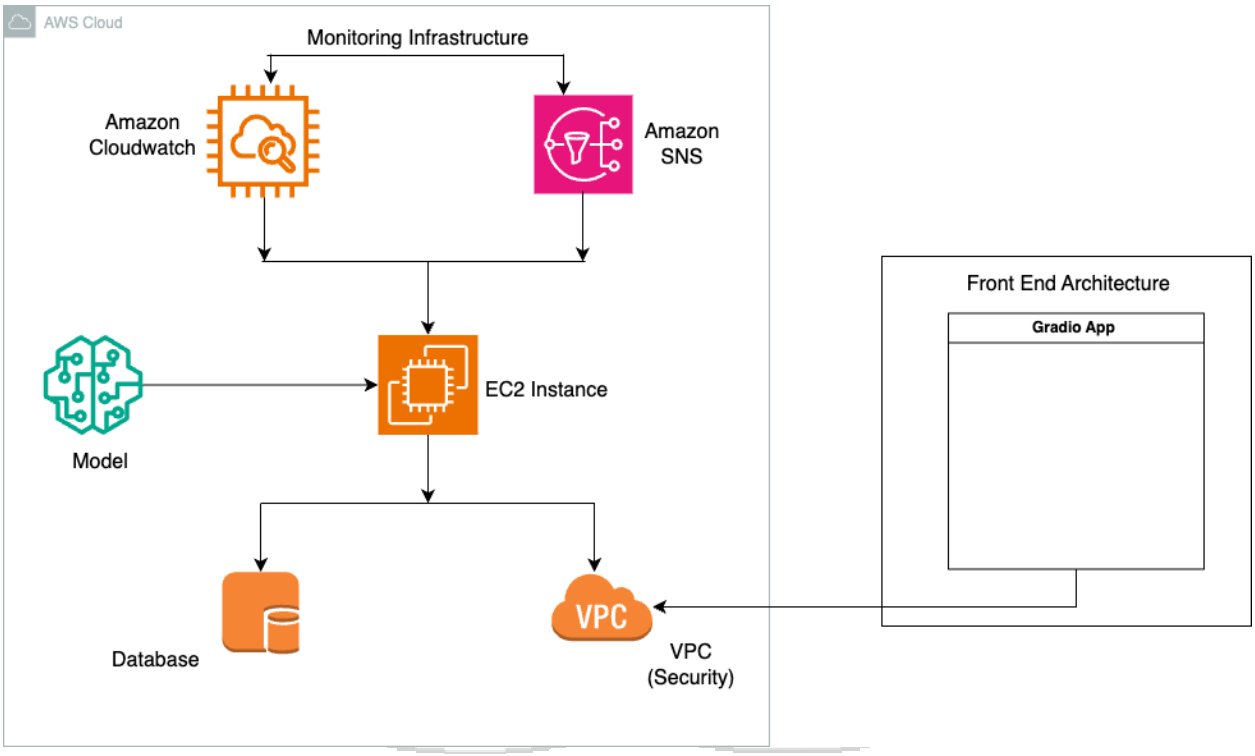


Figure 4.10: Deployment Infrastructure

At the core of this architecture is an Amazon EC2 (Elastic Compute Cloud) instance, which hosts the application logic and loads the pickled model into memory for inference. This EC2 instance is embedded within a Virtual Private Cloud (VPC) to enforce network-level security and restrict unauthorized access. The EC2 environment also manages incoming requests and model execution workflows.

The front-end interface is built using Gradio, a lightweight Python library for deploying machine learning models with interactive UIs. The Gradio app is hosted directly on the EC2 instance, and once the application is launched, the EC2 instance generates a unique URL. This URL acts as a live endpoint that allows users or external systems to interact with the model. Critically, this URL can be embedded into existing web applications or connected to other services via HTTP, making the deployed model accessible for real-time predictions and integrations without requiring additional frontend infrastructure.

The EC2-hosted application also interfaces with a backend database, which is used to persist user inputs, model predictions, and relevant metadata. This database supports traceability and potential model retraining workflows.

To support operational reliability, the deployment is integrated with monitoring infrastructure comprising Amazon CloudWatch and Amazon SNS (Simple Notification Service). CloudWatch is responsible for real-time logging and metric collection, while SNS delivers automated notifications and alerts for anomalies, failures, or threshold breaches in system behavior.

In summary, the model deployment architecture supports end-to-end functionality from secure model serving and user access to observability and integration. The ability to generate an embeddable URL directly from the EC2-hosted Gradio application enhances accessibility and facilitates seamless embedding into broader application ecosystems.

5.0 Conclusion

This study demonstrates the potential of leveraging telematics-based driver behavior data to enhance risk assessment in fleet insurance, independent of traditional claims histories. Through empirical analysis, significant correlations were observed between specific driving behaviors—such as trip frequency, speed, and RPM—and elevated risk levels. Notably, higher frequencies of trips and excessive speeds were found to correlate with increased braking demands and RPM spikes, contributing directly to higher risk scores. These findings underscore the value of using real-time driving metrics as predictive indicators of driver risk profiles. By moving beyond reactive risk models reliant on accident records, this approach offers a proactive, data-driven framework for understanding and managing driver behavior in the context of usage-based insurance (UBI).

Several limitations were noted in the course of this study. The absence of insurer-provided accident records, driver demographic details, and claim histories limited the scope of risk contextualization. Additionally, the presence of non-linear relationships among variables suggests the need for more complex modeling techniques. Future models could benefit from ensemble or hybrid approaches, while algorithmic techniques such as early stopping may be useful in reducing overfitting and improving generalizability.

The findings of this study carry important implications for industry, policy, and research. For insurance providers, telematics-based models facilitate more accurate and behavior-sensitive premium pricing mechanisms. Fleet operators may utilize insights from driver behavior scores to inform risk mitigation strategies and improve operational safety. Policymakers and regulatory bodies, particularly in Kenya and the broader Sub-Saharan region, can leverage these findings to support

structured adoption of UBI frameworks, create national data privacy protocols, and promote the ethical handling of telematics data.

In conclusion, the insights and methodologies presented in this study offer a scalable and impactful foundation for improving safety, efficiency, and policy design in Kenya's transport insurance ecosystem and beyond. The integration of telematics with machine learning has demonstrated its viability as a modern solution to risk profiling challenges in the fleet management sector.

6.0 Recommendations and Future Work

Based on the findings and limitations of this study, several directions for future work and application are recommended. First, integrating insurer-generated data—such as claims history, accident severity, and driver demographic profiles—could substantially improve the accuracy, fairness, and contextual depth of future models. Due to real-world constraints around data sharing, it is recommended that future research incorporate privacy-preserving approaches such as federated learning, which would allow multiple stakeholders to collaborate on training robust models without compromising the confidentiality of their data.

A second recommendation involves the integration of predictive maintenance into telematics systems. Machine learning models such as XGBoost and random forest could be deployed on vehicle edge devices to monitor sensor data in real time, detecting potential mechanical issues before failure occurs. Such deployment would enhance both safety and operational continuity. Transfer learning could also be used to apply trained models across different vehicle types, thus reducing model retraining costs.

Additionally, machine learning can be applied to optimize fleet operations by analyzing telematics data for patterns in fuel consumption, route planning, and driver behavior. Multi-objective optimization techniques can be employed to balance trade-offs between cost efficiency, maintenance scheduling, and delivery timelines. This would significantly improve fleet-wide productivity and resource management.

The integration of telematics with Vehicle-to-Everything (V2X) communication systems is also recommended. V2X-enabled platforms can use real-time data such as speed, direction, and prox-

imity to predict and mitigate collision risks. Furthermore, cybersecurity threats to these platforms can be identified and managed using intrusion detection models trained on machine learning algorithms.

Another important recommendation is the use of telematics as a real-time decision support system. Predictive models can assist drivers and fleet managers with accident risk forecasts, route safety assessments, and dynamic vehicle allocation. These tools could also be used to provide immediate feedback to drivers to encourage safer and more efficient driving practices (Chen and Guestrin, 2016).

Finally, as autonomous vehicles become more prevalent, telematics will play a vital role in enhancing operational intelligence and legal accountability. Sensor fusion involving telematics, LiDAR, and vision data can improve decision-making in autonomous navigation. Telematics data can also support post-incident analysis and system learning in self-driving fleets (Breiman, 2001).

These recommendations reflect the broad potential for advancing the role of telematics in insurance, fleet management, and intelligent transportation. As technology evolves, telematics will remain a key enabler of safer, smarter, and more efficient mobility ecosystems.

References

Aggarwal, C. C. (2017). *Outlier analysis*. Springer, 2nd edition.

Alamir, E., Urgessa, T., GopiKrishna, T., and V, E. (2020). Application of machine learning with big data analytics in the insurance industry. *International Journal of Advanced Research in Engineering and Technology*, 11(12):1064–1073.

Arumugam, S. and Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data. *Journal of Big Data*, 6(1).

Ayuso, M., Guillen, M., and Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. *Transportation*, 46(3):735–752.

Ayuso, M., Guillen, M., and Pérez-Marín, A. M. (2016a). Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4(2):10.

Ayuso, M., Guillén, M., and Pérez Marín, A. M. (2016b). Using gps data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Accident Analysis & Prevention*, 68:160–167.

Boucher, J.-P., Perez-Marín, A. M., and Santolino, M. (2013). Pay-as-you-drive insurance: the effect of the kilometers on the risk of an accident. In *Anales del Instituto de Actuarios Españoles*, volume 19, pages 135–154. Instituto de Actuarios Españoles.

Boucher, J.-P. and Turcotte, R. (2020). A longitudinal analysis of the impact of distance-driven on the probability of car accidents. *Risks*, 8(3):91.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Dasu, T. and Johnson, T. (2003). *Exploratory data mining and data cleaning*. John Wiley & Sons.

Denuit, M. and Trufin, J. (2019). *Effective statistical learning methods for actuaries*. Springer.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.

Finger, R. J. (1988). Risk classification. *Casualty Actuarial Society Forum*. Retrieved September 28, 2023.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks*, pages 1322–1328.

Huang, Y. and Meng, S. (2019). Automobile insurance classification rate making based on telematics driving data. *Decision Support Systems*, 127:113156.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Macedo, L. (2009). The role of the underwriter in insurance. Primer series on insurance, World Bank.

Mungai, D. and Odhiambo, F. (2021). The impact of telematics on fleet safety in nairobi's matatu sector. *Journal of Transport Policy in Africa*, 9(2):112–125.

Mwithimbu, J. (2024). *Matatu madness: A story from the inside*. Eureka Publisher.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons.B*, 21(1):51–62.

Paefgen, J., Staake, T., and Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, 61:27–40.

Palmer, J. (2016). Uninterrupted data transmission and its impact on loss notification in auto insurance. *Journal of Insurance Technology*, 12(1):45–53.

Salehi, A. R. and Khedmati, M. (2024). A cluster-based smote both-sampling (csbboost) ensemble algorithm for classifying imbalanced data. *Scientific Reports*, 14:5152.

So, J., Kim, A., and Lee, S. (2021). A cost-sensitive multi-class adaboost algorithm for estimating accident frequency using vehicle telematics data. *Decision Support Systems*, 150:112–123.

Sun, J., Kim, S., and Lee, H. (2020). Utilizing average accelerator pedal position and brake count to measure driving risk under limited accident data conditions. *Journal of Transportation Analytics*, 12(3):345–360.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

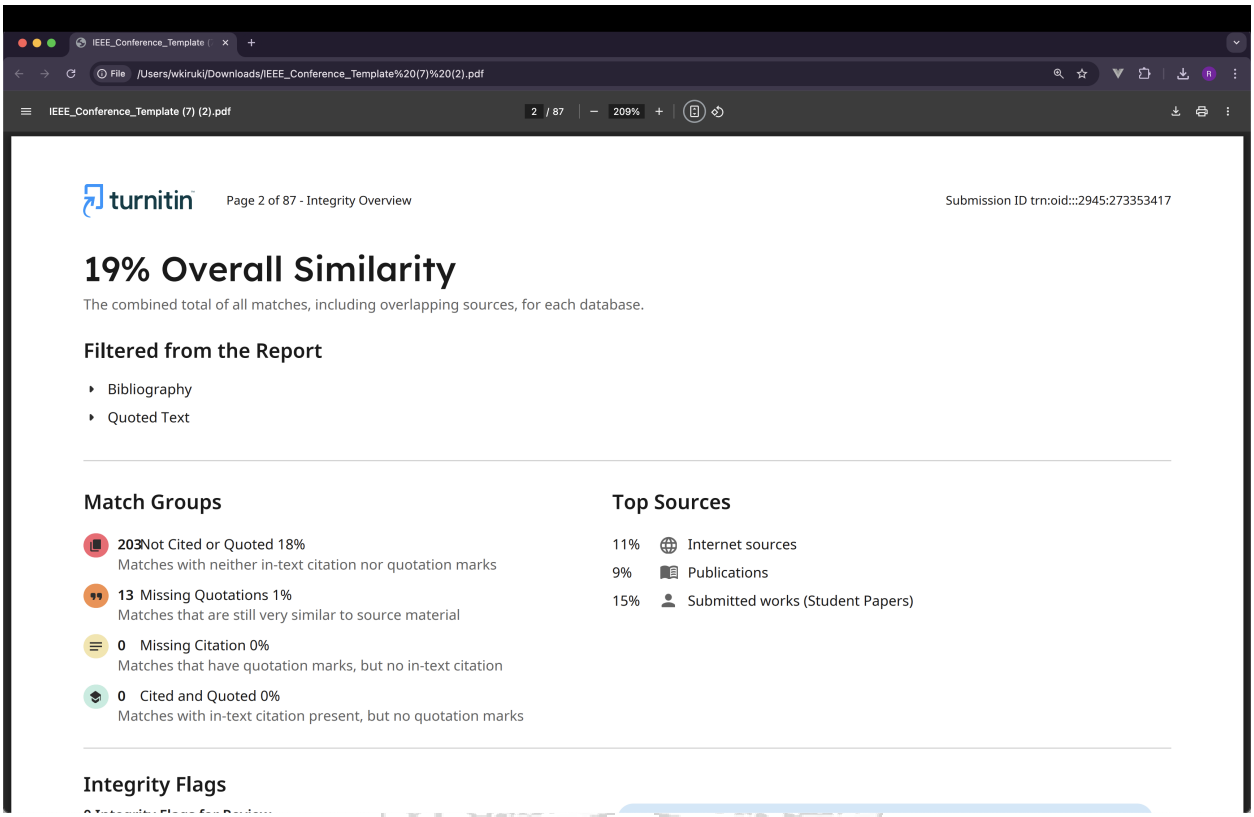
Warren, G. S. and Greenlee, M. R. (2006). Calculation of driver score based on vehicle operation.

Yao, Y. (2018). Evolution of insurance: A telematics-based personal auto insurance study. *Honors Scholar Theses*, 11(590).

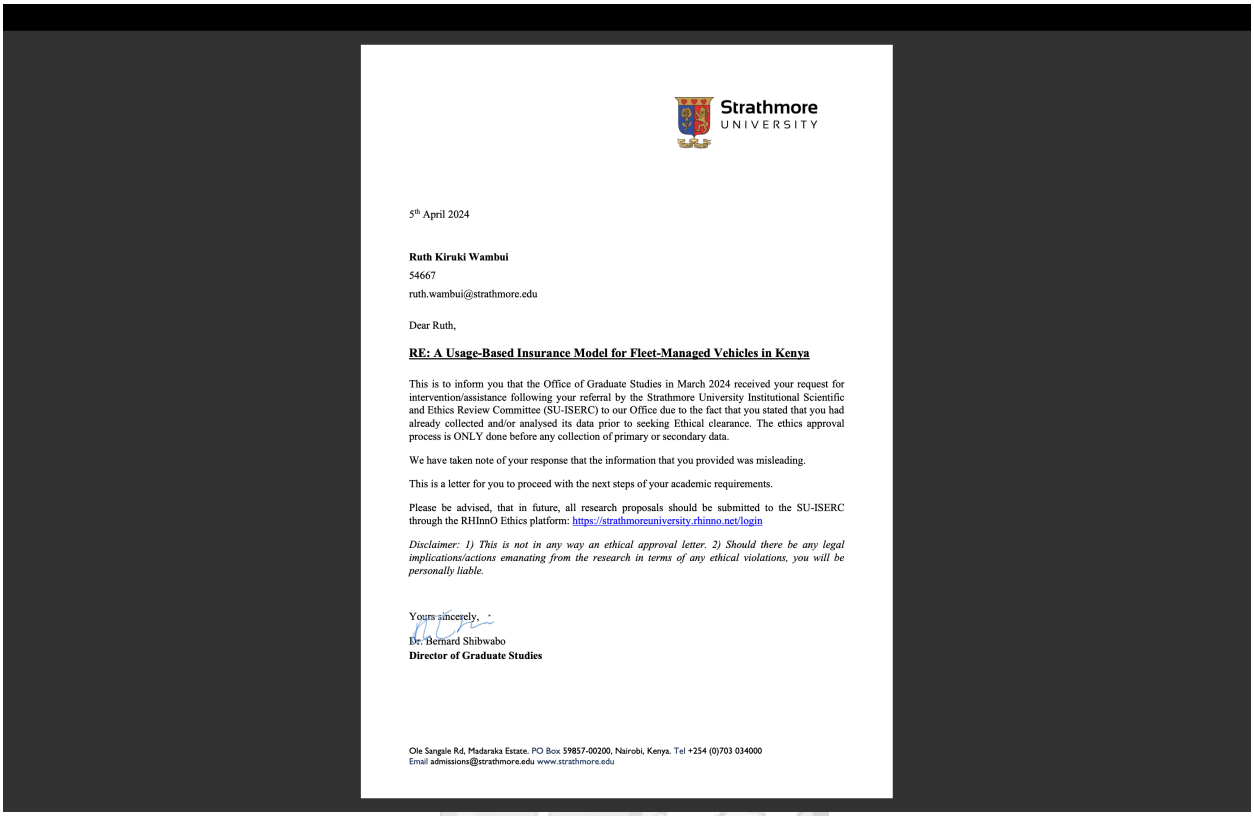
Zhu, T. (2017). Sensor data, privacy and behavioral tracking: Does usage-based auto insurance benefit drivers? Technical report, University of British Columbia.

Appendices

Appendix A: Plagiarism Report



Appendix B: Ethical Clearance Release Letter



Strathmore UNIVERSITY

5th April 2024

Ruth Kiruki Wambui
54667
ruth.wambui@strathmore.edu

Dear Ruth,

RE: A Usage-Based Insurance Model for Fleet-Managed Vehicles in Kenya

This is to inform you that the Office of Graduate Studies in March 2024 received your request for intervention/assistance following your referral by the Strathmore University Institutional Scientific and Ethics Review Committee (SU-ISERC) to our Office due to the fact that you stated that you had already collected and/or analysed its data prior to seeking Ethical clearance. The ethics approval process is ONLY done before any collection of primary or secondary data.

We have taken note of your response that the information that you provided was misleading.

This is a letter for you to proceed with the next steps of your academic requirements.

Please be advised, that in future, all research proposals should be submitted to the SU-ISERC through the RHInnO Ethics platform: https://strathmoreuniversity.rhinno.net/login

Disclaimer: 1) This is not in any way an ethical approval letter. 2) Should there be any legal implications/actions emanating from the research in terms of any ethical violations, you will be personally liable.

Yours sincerely,

Dr. Bernard Shibwabo
Director of Graduate Studies

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix C: Driver Behavior Project Repository

This appendix references the GitHub repository containing code, data, and documentation for the driver behavior analysis project.

i. Repository URL:

   `https://github.com/RuthieKe/Driver-behaviour-project.git`

ii. Clone command:

   `git clone https://github.com/RuthieKe/Driver-behaviour-project.git`

iii. Last accessed: May 29, 2025

The repository contains:

i. Python/Jupyter Notebook code for data analysis

ii. Raw and processed datasets and code for the Gradio app