**RESEARCH METHODOLOGY**

**3.1 Introduction:**

This chapter explains why this particular problem related to fraud detection should be studied.

It describes the research methodology, including each of the components that the researchers regard as reality and the truths related to their underlying research philosophy. In addition, this chapter explains the data management, model split and training, including the research philosophy.

**3.2 Research Philosophy**

When deciding the research philosophy that would suit this study, the stumbling block was primarily understanding the various terminologies and explanations surrounding research philosophies; the difficulty of this has been admitted by studies like (Mkansi M. et al.2012), who noted that the differences in the use and inter-switching of these terms could act as a stumbing block to researchers whom are new to the field of research.

The research philosophy involved in this research is rooted in pragmatism. The study by Allemang B. et al. (2022) gave a brief definition of pragmatism, stating that “Pragmatism as a paradigm is based upon the premise of utilizing the best methods to investigate real-world problems, allowing for the use of multiple sources of data and knowledge to answer research questions.”

The researcher focus on solving a tangible issue – fraud detection – by using adaptive machine learning techniques that responds dynamically to changes in transaction data. The use of online learning is crucial to maintaining the model’s adaptability, which does reflect a paragmatic approach that prioritizes real-time response capabilities over static one-time solutions.

Fraudlent behaviour is dynamic, requiring a nuanced approach that does consider both algorithmic results and expert evaluations to make decisions. Therefore the combination of computational techniques with interpretive analysis shows the importance of a research philosophy that is both flexible and outcome-oriented, and aligned with the complex nature of fraud detection in financial systems.

**3.3 Research Sampling**

The researcher adopted a sampling strategy that has a focus on ensuring a representative and balanced dataset to train and test the models effectively. Given the true nature of fraud detection, the dataset typically represents a significant observation of class imbalance, with fraudulent transactions being vastly outnumbered by the legitimate transactions.

The dataset that the researcher used for this research consists of historical transaction records, which was sampled to reflect closely to a wide variety of scenarios that are likely to occur in a real-world financial system.

To address the issue of class imbalance, the researcher employed the technique of oversampling the minority class using methods such as Synthetic Minority Over-Sampling Technique (SMOTE) to augment the dataset during the model training. This approach helps to balance the class and prevents the model from becoming skewed towards predicting only the majority class. Techniques such as SMOTE have been widely validated in fraud detection research as an effective means to fix the issue of data imbalance

**3.4 Ethical Considerations**

The researcher acknowledged that given the sensitivity of the financial data involved, strict data anonymization techniques were applied to remove any personally identifiable information (PII) before data processing, thereby protecting the privacy of individuals. Why anonymization was used is because it ensures that no user can be identified directly or indirectly from the dataset, which is aligned with data protection regulations such as the General Data Protection Regulation (GDPR).

These laws were chosen as a placeholder to show that no intellectual properties or digital identities of the information provided by the dataset would be used outside the scope of this study.

**3.5 Accessibility, Reliability and Validity**

This project in its entirety would be made available at all times for the use of students and other researchers, which would aid in making it as accessible within the reach of any researcher. The dataset used, being a public dataset, is also available alongside the notebook containing the various models and their results from training and testing. The robustness of the constructed models, the use of SMOTE and cross-validation, and the data quality all contributes to the study’s reliability. To help remove any potential inaccuracies originating from the dataset, the researchers subjected it to a rigorous and in-depth pre-processing approach.

Validity refers to the model’s ability to measure what it is intended to – effectively distinguish fraudulent transactions from a legitimate transaction. Internal validity was done by the process of data pre-processing, such as handling class imbalances and employing techniques such as SMOTE to ensure the model’s outcomes are reflective of its ability to detect a fraudulent transaction. External validity, or the ability to generalize findings beyond the specific dataset used is also prioritized.

**3.6 Data Collection, Handling and Implementation**

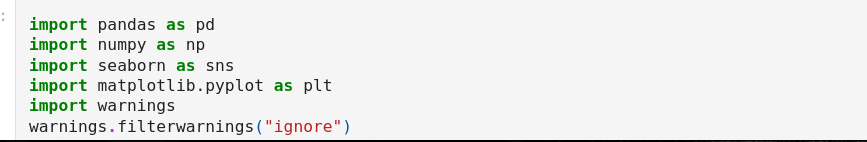
**3.6.1 About the Dataset**

The dataset used in this project consists of 151112 rows and 11 columns.

It contains data such as the signup\_time, purchase\_time, purchase\_value, browser, age.

**3.6.2 Dataset Loading and Preprocessing**

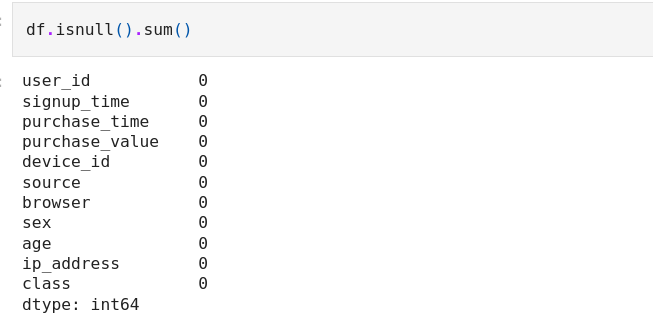
Python imports a data analysis library. This library provides the needed data structures like DataFrame and Series, which are crucial for handling structured data.



A code snippet showing the package import for data preprocessing (Created by the researcher)

The dataset used in this research was loaded using the pandas library, which offers data handling and manipulation capabilities, which are needed for dealing with large financial transaction datasets. The data which was originally stored in a CSV file, was read into a DataFrame (df) to facilitate both data exploration and preprocessing. The initial data analysis focused on understanding the dataset’s structure which includes the data types, distribution of values, and identifying missing values using both df.info() and df.isnull().sum(). This step was crucial for assessing the quality of the data for further model development.

Data preprocessing involved multiple stages to prepare the dataset for effective model training. First, missing values were addressed, although it was found that no missing values existed in the dataset, this made way for a direct progression to the next steps.

Code snippet showing the observation of no missing values (Created by the researcher)

The exploration of the dataset by the researcher showed a significant imbalance between the number of fraudulent (fraud\_label) and non-fraudulent (non\_fraud label).

Since fraud detection involves a case of highly imbalanced classes, with non-fraudulent transactions outnumbering fraudulent transactions, a resampling technique called Synthetic Minority Over-Sampling Technique (SMOTE) was applied by the researcher. SMOTE helped to balance the dataset by generating synthetic examples of fraudulent transactions, this makes sure that the machine learning models the researcher will build could learn to identify fraud effectively, rather than becoming biased towards the majority non-fraud class.

**3.7 Limitation**

One major limitation of this research is the presence of class imbalance in the dataset, where the number of non-fraudulent transactions exceeds that of fraudulent transactions. Presence of this imbalance poses a challenge for model development and training, as the model algorithms could be biased towards the majority class, and this could lead to a poor performance in detecting any fraudulent transaction. Techniques such as SMOTE which was recommended by the researcher could introduce a syntethic noise into the dataset, and this could potentially reduce the model reliability.

Another limitation is the dependency on historical data, which may not fully capture emerging fraudulent transactions patterns that could evolve in real-time.

The use of static datasets, even with adaptive online learning, can result in models that are less effective at responding to a different and new sophisticated fraud schemes that have not been seen before.

The reliance on deep learning techniques, although it’s beneficial for capturing complex relationships in data, introduces challenges regarding model interpretability. Financial institutions and stakeholders often require a very clear understanding of why any transaction should be flagged as fraudulent to maintain trust and comply with regulatory requirements. Due to the “black box” nature of deep learning models, this complicates the issue of providing transparent and actionable insights.