# Chapter 3: Methodology

## 3.1 Introduction

The reliability of any research is greatly determined by the quality of the methods used in that research. Further details of the research design and method used in this study which involves both machine learning and survey analysis for customer value prediction and assessment of the likelihood of customers’ adoption of a unified membership system are detailed in this chapter. Python-based tools are used to put into practice these methods to make the tasks accurate and easily reproducible. This chapter presents a clear understanding of the research philosophy, the approach used and the research methods.

## 3.2 Method Outline

The research follows basic steps of data gathering, data cleaning, data transformation, model construction, model assessment and survey data analysis. All of the project phases employ state-of-the-art machine learning and stats tools integrated into Python packages including pandas, NumPy, scikit-learn, XGBoost, and Optuna. Data from transaction and survey provides a solid platform where a thorough analysis of the customer behaviours and their preferences can be made. Loading and cleaning of the datasets mark the first stages of the data pipeline process. Feature engineering, converts raw data into final and useful, labels such as Recency, Frequency, Monetary (RFM) scores. Customer life time value predictions models are then built in order to estimate customer lifetime value (CLV). As a result of performing clustering techniques for example KMeans, it is easy to segment customers for actionable insights view on customers.

## 3.3 Research Philosophy

The research is based on the positivism research paradigm and supports the idea that knowledge is based on observation and facts. This philosophy is well-suited in the course of the study to estimate the customer value and behaviour and is based on quantitative figures (Adekoya and Aneiba, 2024). Another method of machine learning models is free from the subjective interpretation and presents actual data, thus, following the principles of positivism, which stems from the need for accurate and replicable data proof.

## 3.4 Research Approach

This research follows the deductive research approach where the research hypotheses are developed based on previous theories and models of customer behaviour and the adoption of the loyalty program. These assumptions are generated from the literature and the models are used to test these hypotheses. The deductive approach makes certain that the research is grounded with theory.

## 3.5 Research Design

The study adopts a quantitative and exploratory research approach. Customers are analyzed and segmented using quantitative methods to forecast their behaviors about buying and responding to questionnaires (Zhou *et al.* 2024). The nature of the current study can also be seen from the fact that the clustering analysis aims at identifying hidden characteristics within the data. The split approach increases the robustness of the model since it includes not only precise predictions but also exploratory analysis of customer segments. Due to the work’s focus on the predictive model, the high degree of machine learning is needed, namely XGBoost and deep neural networks, to improve the predictive accuracy in addition to the model’s interpretability.

## 3.6 Research Method

Quantitative method is used to analyze the transactional data while the qualitative method, survey responses are used to gain insights. The numerical data is managed by machine learning models, but the survey data is used to offer another layer of insight into customer preferences and likelihood of adoption (Reddy *et al.* 2024). This integration allows for a deeper understanding of customer performance while incorporating both the frequency and the nature of performance. This mixed method was made possible by the Python ecosystem that leveraged throughout his research (Yang *et al.* 2022). For example, survey data preprocessing consisted of discrete feature encoding and numerical feature scaling in order to make the data compatible with transactional data and integrate it.

## 3.7 Data Collection Method

To cover all aspects of customer behavior and preferences, data collection for the study was designed in two streams. The first stream, transactional data was retail transactions recorded over a period of time. This dataset means it was able to record customer purchases in greater detail including issues like quantity, price and date of purchase. The second type of data was collected through the survey data that was taken from the survey results. The following survey data include basic fields of the customers, their attitude and preference to the newly proposed integrated membership scheme (Nele *et al.* 2024). Special attention was paid to the selection of high quality and topical sets of indicators. Before the actual analysis, the first and second data stream were preprocessed and imported into Python.

## 3.8 Data Preprocessing and Feature Engineering

Cleaning and preparation of the data from the datasets followed rigorous procedures so that the final data was raw and suitable for analysis. The first step involved dealing with missing values and duplicate records as well as outliers which may distort the outcome. Some of the preprocessing steps performed include data cleaning where rows with missing or invalid customer ID where deleted. Both quantitative features were next standardized using scikit-learn’s StandardScaler, bringing all numerical typed features to the same measurement scale, which is crucial for many machine learning models (Neelakandan *et al.* 2023). Feature engineering was critical throughout because of the improved predictive capabilities of the given dataset. New variables were introduced to reflect other important characteristics of customer behavior. For instance, the Total\_Amount feature was derived through simple arithmetic of multiplying price and quantity giving an actual metric of customers’ expenditure. RFM scores with recency, frequency, and monetary values were obtained for capturing the behavior of the customers since the model needed more clarity regarding the customers’ engagement. Time characteristics including the purchase day of the week, month, and time were derived to capture temporal patterns regarding customer behavior. All these engineered features greatly helped the model to predict the actions of a customer more aptly enhancing the performance of all the machine learning model.

## 3.9 Model Development

The study used two main predictive methods which include XGBoost and an artificial neural network. Among the classification algorithms, XGBoost, a gradient-boosting algorithm, was chosen because of higher predictive performance and interpretation of the results using feature importance for the structure data (Peng and Unluer, 2024). To further improve the efficiency of this model, hyperparameter tuning by Optuna was performed to increase the accuracy with the right selection of parameters such as max\_depth, learning\_rate, and n\_estimators. This tuning process also made it possible for the model to attain the best balanced accuracy at the validation dataset. However, a neural network was also created to model more complex linear relationships in the data set. Training stopping criteria were used to stop the training process when the validation loss was not minimizing any longer, in order to save on unnecessary epochs of training. As for the performance of models, several baseline models including RMSE and MAE predicted the model accuracy in detail, and the R2 measure quantified the variance of each model. The results of the study were also supported by simple visual confirmation such as scatter plots of the actual vs. the predicted values from the models.

## 3.10 Research Ethics

In the Research process, ethical issues were accorded paramount significance especially when dealing with customers’ information. In order to avoid compromising privacy and adhere to ethical norms some important steps were taken throughout the process of research. The first measure was data anonymization, which entailed either the stripping or partial concealment of identity information (Arunachalam and Kumareshan, 2024). This approach was important in reducing the danger of giving personal information which can be leaked or misused. First of all, the ethical consideration of informed consent was a part of the study. Every respondent in the survey was adequately informed of the objectives of the study, collection of their information and its usage. Such clarity made it possible for participants to volunteer and provide their consent willingly knowing the fact they had the right to pull out from the study at any one time. One was on the data security aspect which was deemed to be very critical. Safeguard measures to contain the collected data were adopted so as to prevent it from been accessed or hacked by third parties.

## 3.11 Research Limitations

This research faced some challenges that affected the general conduct of the research study. Data quality was one of the main issues raised, which meant that inconsistent entries were made either during the analysis or towards the beginning of the project, thus skewing the results in some way. Moreover, the generalization ability of the developed models was quite restricted as these were developed specifically for the current dataset which hamper its use in other domain or other scenarios (El-Shorbagy *et al.* 2024). Other limitations included computational capacity since the manner in which the models were optimized, especially through hyperparameters, was limited by available computational resources.

## 3.12 Conclusion

This chapter amplified a comprehensive discussion of the method used in the research by especially focusing on the data acquisition, cleaning and preprocessing, feature extraction and transformation, model training, and assessment. It focused the effectiveness of Python’s superb library in terms of processing and manipulation of raw data into valuable information. Data preprocessing and data cleaning which helped in removing the irrelevant data, and this helped in producing accurate results Feature engineering also improved the feature set in the given data set and improved the variables reflecting the customer behaviors and preferences. Using data from transactional databases and survey, the approach facilitated closing the gap between the quantitative and qualitative perspectives. Customers and their adoption likelihood of a product were predicted using machine learning models like XGBoost and neural network. The study went a step further to fine tune the hyperparameters and test the accuracy of the developed models. Safety of participants’ data privacy and their consent was respected thus following the best practices in data protection.