**Title of Dissertation**

# Data-Driven Profitability: Optimizing Credit Card Business through AI-powered Customer Retention & Risk Management

**Module Code: BUSN9860**

**Module Title: Dissertation**

**Advisor: Zhen Zhu**

**Programme of Study: MSc Business Analytics**

**Surname: Sen**

**First Name: Tuhin**

**Login: ts703**

**Required Word Count: 8,000 – 10,000**

**Actual Word Count: 10,929**

I agree that copies of my report may be used as reference material by the University (please tick)

Yes [✓] No [ ]

Ethics (please tick):

[ ] **I HAVE** received ethical approval from the REAG Chair before conducting my research

[ ] **I HAVE NOT** received ethical approval from the REAG Chair before conducting my research

[✓] My research did not involve human participants so I was not require to receive ethical approval

# Acknowledgement

I'd like to express my heartfelt appreciation to my supervisor, Professor Zhen Zhu for his exceptional support, and encouragement throughout this journey. His unwavering patience, extensive knowledge, and invaluable guidance while allowing me the freedom to explore my ideas has been truly empowering. This journey would not have been as enriching and fulfilling without his mentorship.

I am grateful to my friends, and family, who have emotionally supported me throughout this journey all the way from the beginning. Their encouragement, understanding, and faith in me have been a constant source of motivation, helping me persevere through the challenges and celebrate the milestones along the way.

Lastly, I would like to thank my course instructors at the University of Kent for providing me with the skills and theoretical knowledge I needed to conduct this research. Their dedication to teaching and commitment to student success have provided a solid foundation for my academic and professional development.

Thank you for your invaluable contributions to my academic journey. This thesis could not have been completed without your guidance, support, and belief in my abilities.

# Table of Contents


Chapter 1: Introduction ..... 5
Chapter 2: Literature Review ..... 6
2.1. Literature 1: ..... 6
2.2. Literature 2: ..... 7
2.3. Literature 3: ..... 8
2.4. Literature 4: ..... 8
2.5. Literature 5: ..... 9
Chapter 3: Methodology & Data ..... 10
3.1. Literature Review Methodology: ..... 10
3.2. Approach To Analysis Overview: ..... 10
3.2.1. Data Modelling Process: ..... 11
3.2.2. Data Clustering Process: ..... 12
3.3. Feature Engineering Techniques: ..... 12
3.3.1. Data Standardisation: ..... 12
3.3.2. SMOTE (Synthetic Minority Oversampling Technique): ..... 12
3.3.3. PCA (Principal Component Analysis): ..... 13
3.4. Machine Learning Algorithms: ..... 13
3.4.1. Logistic Regression: ..... 13
3.4.2. Decision Tree Classifier: ..... 14
3.4.3. Random Forest Classifier: ..... 14
3.4.4. Diagrammatic Example of Bagging Classification (Random Forest) ..... 15
3.4.5. Extreme Gradient Boosting Classifier: ..... 15
3.4.6. Light Gradient Boosting Classifier: ..... 16
3.4.7. Diagrammatic Example of Boosting Classification (Boosting Algorithms): ..... 16
3.5. Clustering Algorithms: ..... 17
3.5.1 K-means Clustering: ..... 17
3.5.2. Hierarchical Clustering: ..... 17
3.6. Evaluation metrices: ..... 18
3.6.1. Precision: ..... 18
3.6.2. Accuracy: ..... 18
3.6.3. Recall: ..... 19
3.6.4. F1: ..... 19
3.6.5. Davies Boulding Index: ..... 19

3.6.6. Silhouette Score: 20
3.6.7. Calinski-Harabasz Index: 20
3.8. Data Description: 20
3.8.1. Class Imbalance: 21
Chapter 4: Analysis & Insights 21
4.1. Clustering Analysis: 22
4.1.1. Credit Card Customer Churn: 22
4.1.2. Credit Card Default: 26
4.2. Modelling Analysis & Comparison: 31
4.1.1. Credit Card Customer Churn: 31
4.2.2. Credit Card Default: 32
Chapter 5: Recommendations & Discussion 33
5.1. Strategies and Personalized offers (Research Question 1): 34
5.1.1. Credit Card Customer Churn: 34
5.1.2. Credit Card Default: 35
5.1.3. Key Financial Behaviour from Customers in Both Datasets: 36
5.2. Modelling Results & Business Impact Discussion (Research Questions 2): 37
5.2.1. Business Impact of LightGBM Model for Churning: 37
5.2.2. Business Impact of Random Forest Model for Defaulting: 39
5.3. Model Interpretability & Feature Importance (Research Question 3): 42
5.3.1. LightGBM for Customer Churn: 42
5.3.2. Random Forest for Credit Card Default: 44
5.4. Future Research Directions: 45
Chapter 6: Conclusion 46
Appendices 48
Appendix 1: 48
Appendix 2: 49
Appendix 3: 55
Appendix 4: 57
References 60

# Abstract

This study investigates the transformative potential of Artificial Intelligence (AI) and Machine Learning (ML) in the credit card industry for reducing customer churn and credit defaults, increasing profitability and risk management. The study provides actionable insights for improving the customer experience and financial outcomes by utilizing advanced data analytics and Explainable AI (XAI) techniques.

The study uses clustering analysis with the K-means algorithm to identify key customer behaviours, revealing seven distinct groups: budget-conscious families, loyal long-term customers, high-net-worth customers, credit-reliant rising spenders, financially savvy spenders, financially strapped payers, and risk-taking credit consumers. High credit utilization and financial strain are common behaviours, indicating the need for integrated retention and risk management strategies.

A comparative analysis of various machine learning models balances accuracy and recall when predicting churn and defaults. The Light Gradient Boosting Machine (LightGBM) for churn prediction achieved 96% accuracy and 88% recall, reducing churn from 16.07% to 11.82% and saving approximately NT$430,000 based on the assumption that the average revenue per customer is NT$1000. The Random Forest Classifier reduced default rates from 22.11% to 18.27%, resulting in a 51% recall and a savings of approximately NT$25,000 based on the assumption that the average loss per default is NT$1000.

XAI techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) were used to identify key drivers of churn and default. SHAP analysis focused on transaction frequency and amounts, whereas LIME identified missed payments and education levels as significant predictors.

# Chapter 1: Introduction

The credit card industry faces ongoing challenges in balancing customer loyalty and credit risk (Rane, 2023). Historically, this balance was determined by static credit scores and generic marketing campaigns. However, in today's competitive landscape, which is characterized by high customer churn due to appealing offers from competitors, effective customer retention strategies are essential (Rane, 2023). Churn reduces profitability due to lost revenue and the high cost of acquiring new customers. Concurrently, credit risk, defined as defaults when borrowers fail to meet payment obligations, causes significant financial losses for credit card companies. Economic downturns, unexpected financial difficulties, and reckless credit use amplify this risk.
This is where AI and machine learning (ML) come into the picture in transforming the credit card business (Sathvika et al., 2024). By leveraging AI and ML businesses can personalizes offers and predicts churn with high accuracy, increasing loyalty and preventing customer loss (Bharadiya, 2023). AI can also be used to improve credit risk assessment by analysing large datasets to detect subtle patterns that predict defaults. This allows for more strategic credit extension, reducing risk while increasing profitability.

This research leverages AI and ML to develop strategies for customer retention and credit risk mitigation strategies. Initially, datasets are prepared for clustering and modelling using data transformation and cleansing techniques. Clustering analysis is employed using K-means and agglomerative clustering to decide the optimal number of clusters (Miraftabzadeh et al., 2023). This evaluation is done using Davies-Bouldin Index, Silhouette Score, and Calinski-Harabasz Index. Further on the optimal clusters are explored to discover insights. Next the datasets are prepared for modelling through using feature engineering techniques such as SMOTE and standardisation for tackling class imbalance in the target variable (Abedin et al., 2023). The selected ML algorithms include Logistic Regression, Decision Trees, Random Forest, Extreme Gradient & Light Gradient Boosting Classifiers, are chosen for their interpretability (Chen, Calabrese, and Martin-Barragan, 2024). The model evaluations are conducted on 30% of the unseen dataset to assess model generalizability. Feature importance is interpreted for the champion models in both credit card churn and default cases by using XAI techniques (Bello, 2023). Model comparisons are based on Recall, Accuracy, Precision, and F1 score. Finally, the study concludes with discussion on the insights uncovered from the analysis and further research for improvement.

This review investigates the potential of Artificial Intelligence (AI) and data analytics to maximize profitability and minimize risk of the credit card business by mitigating defaulting and customer churn. The primary objectives of this study are:

1. How data analytics can be utilized to identify key customer behaviours and develop targeted offers, retention programs, and risk management strategies that maximize profit? Are there any common customer behaviours from both sets of data?

2. How can comparing various machine learning models help to find the best balance of accuracy and recall, thus identifying the potential impact of missed churners and defaulters? Are there different best performing models for each task – churning prediction and defaulting prediction? If so, then what are the potential financial impacts of the best models for each task?

3. What XAI techniques can be employed to identify the key drivers of credit card default and customer churn?

The following provides the structure of this paper. 1. A comprehensive review of existing literatures relevant to the topic, by highlighting the strengths, weaknesses and potential gaps in the research. 2. Then, it explains the dataset and methodology in depth. 3. Furthermore, the study then provides a detailed discussion on the insights extracted from the analysis and 4. concludes this research with strategic recommendations to improve the credit card business in terms of customer retention and credit risk mitigation.

# Chapter 2: Literature Review

## 2.1. Literature 1:

This research by (Lleberi, Sun and Wang, 2024) aims to predict credit risk using ML modelling techniques. The approach presented in the paper is more directly applicable to customer churn, even though it is positioned within the context of credit risk assessment. It uses three different datasets namely the German, Australian and Taiwan credit datasets from UCI. The author utilizes multiple ML algorithms to assess model performance where the stacked ensemble model outperforms individual machine learning algorithms (RF, GB, XGB, KNN, ANN, and DT) in terms of credit risk prediction. Strong credit risk prediction assists in identifying customers who are more likely to default on their credit card payments (Xie et al., 2021). Banks can significantly reduce their financial losses from defaults and write-offs by not issuing cards to these high-risk individuals. This increases the overall profitability of their credit card business.

One of the key strengths of this paper is that the research uses an Information Gain (IG)-based feature selection method to identify the most important factors influencing customer risk (Bouchlaghem, Akhiat and Amjad, 2022). This can improve model interpretability while potentially reducing the resources required for model training. This study also compares the stacked model to previous research on the same datasets, demonstrating its model's prediction performance. This increases the credibility of the proposed approach.

While the paper mentions improved model performance, it does not explicitly link these findings to tangible business benefits. A more thorough analysis could quantify the potential revenue gains for banks from mitigating risk or avoiding handing out credit cards to clients with history of default. The paper also does not discuss how class imbalance is handled in those datasets. There are also some gaps in the research which could have been explored such as the study does not address recommending specific products or services to have an impact on the business. Data analytics can be used to analyse customer data and recommend relevant credit card products or features (e.g., rewards programs, balance transfers) that are more likely to appeal to each customer segment or individual. The study could have also conducted an exploratory data analysis to give readers a baseline understanding of the distributions of each dataset. Each dataset could also have been further utilized to form clusters from each, and those

clusters could have been used in modelling to get better understanding of credit risk prediction factors.

In conclusion, (Lleberi, Sun, and Wang, 2024) present a strong machine learning approach to credit risk prediction that employs a stacked ensemble model to achieve significant accuracy gains. Their use of multiple datasets and Information Gain-based feature selection improves model interpretability and risk identification, which is critical for reducing financial losses due to defaults. However, the study does not establish a direct link between improved model performance and tangible business benefits, such as quantifying revenue gains or addressing class imbalance as doing so could provide more clarity about the model performance. Furthermore, investigating product recommendations and conducting exploratory data analysis may increase the study's practical relevance and applicability.

## 2.2. Literature 2:

Another unique study by (Jovanovic et al., 2024) examines the integration of blockchain and explainable artificial intelligence for credit scoring. The primary goal of the reviewed literature was to investigate how blockchain and federated learning (FL) technologies could be used to improve credit assessment models. This includes improving model verification, reliability, transparency, and explainability—all of which are critical for optimizing credit card business strategies aimed at customer retention and risk management (Olateju et al., 2024).

One of the strengths mentioned in the literature is the use of blockchain to ensure data integrity and security. Blockchain protects credit scoring decisions by immutably recording model updates and parameters (Yang, Abedin, and Hajek, 2024). This is critical for maintaining customer relationships and reducing default risks, particularly in credit card companies where data privacy is paramount, thereby protecting customer privacy and promoting regulatory compliance. This enables more accurate risk assessments with diverse datasets. These strengths are critical for credit card companies to maintain customer trust by handling data securely and providing personalized services.

Despite these strengths, several weaknesses were discovered. One major challenge is the scalability of blockchain solutions in federated learning frameworks (Asqah and Moulahi, 2024). Blockchain's computational and storage requirements may limit its scalability when dealing with large credit card transaction and customer profile datasets, affecting real-time risk management and retention strategies. Although FL improves privacy, combining it with blockchain and Explainable AI (XAI) techniques presents implementation challenges (Ullah et al., 2023) Seamless integration and operational efficiency across multiple systems continue to be barriers to adoption. The interpretability factor is another weakness of this study. Even with XAI, stakeholders who need to understand credit scoring decisions continue to be concerned about AI model interpretability (Chamola, et al., 2023). Research into improving AI model interpretability could result in more robust credit scoring systems that can meet regulatory requirements.

While blockchain-enabled federated learning frameworks show promise for optimizing credit card business through improved customer retention and risk management strategies, it is critical to address existing weaknesses and close identified gaps. Credit card companies can improve operational efficiency and customer trust by leveraging strengths like data security, privacy protection, and model transparency (Rane, 2023). However, overcoming challenges such as

scalability, integration complexity, and model interpretability is critical for successful implementation in real-world business settings.

## 2.3. Literature 3:

Another review by (Grodzicki et al., 2023) focuses on the analysis of credit card user behaviour in response to price changes, as well as the implications for policy and financial strategies. The primary goal is to better understand how consumers react to changes in interest rates and fees, and how these responses can be used to inform targeted offers, retention programs, and risk management strategies. This analysis provides insights into consumer decision-making processes that can be used to increase profitability for financial institutions.

According to (Dong & Yang, 2020), one of these studies' major strengths is that they rely on extensive data from a large database of credit card accounts, which adds robustness and depth to the analysis. Another strength is the use of behavioural economics theories, which improves understanding of consumer decision-making processes. This theoretical lens is especially useful for studying subprime users' behaviour, as it highlights how psychological factors influence financial decisions (Lim et al., 2022). Such insights are critical for developing successful financial strategies and policies.

However, the literature has its limitations. The emphasis is primarily on credit card usage, which may not encompass the full range of consumer financial behaviours across various financial products. Furthermore, the exclusion of cardholders who are more than 60 days late, as well as those with newly formed or closed accounts, results in a selection bias (Rajput, Wang, and Chen, 2023). This may underestimate the behaviours of financially distressed users, limiting the findings' generalizability. Another significant limitation is the indirect examination of users' entire balance sheets. This gap limits our understanding of the entire financial context in which consumers make decisions, potentially leading to an incomplete assessment of the factors that influence credit card usage and repayment behaviour (Bello, 2023). Furthermore, while the findings are strong within the study's context, their generalizability to other financial contexts or products may be limited.

Finally, the literature establishes a solid foundation for using data analytics to gain insights into customer behaviour, which can then be used to develop targeted financial strategies (Sheth, Jain and Ambika, 2023). The empirical analysis and application of behavioural economics provide a more nuanced understanding of consumer responses to price changes. However, the limitations, such as the exclusion criteria and the lack of comprehensive transactional data, highlight the need for a broader approach to fully comprehend the diversity of financial behaviours (Adeyeri, 2024). Integrating these insights into data-driven strategies is critical for aligning financial offerings with customer behaviours and regulatory environments, resulting in increased profitability and strategic effectiveness (Oriji et al., 2023).

## 2.4. Literature 4:

A study by (Gangadhar et al., 2023), aims to understand and predict customer churn in B2C e-commerce using machine learning techniques. By analysing client exchange data over a year, the study hopes to identify patterns and attributes that contribute to customer turnover. The

study focuses on data standardization and normalization, as well as the classification and prediction of customer churn using artificial neural networks and support vector machines (SVMs). The goal is to develop strategies for retaining existing customers, increasing profitability, and fostering long-term business growth.

One significant strength of this study is its thorough data analysis. Using a large data set to understand customer behaviour and transactions, resulting in accurate churn predictions. This enables more efficient resource allocation towards retaining valuable customers (Alizadeh et al., 2023). Furthermore, the use of advanced machine learning techniques like artificial neural networks and SVMs is a significant advantage. These deep learning techniques effectively manage complex relationships in customer data, resulting in higher prediction accuracy and lower churn rates (Ahmed et al., 2023). Additional strengths include data standardization and normalization to address errors and improve analysis reliability, resulting in better decision-making (Santos, 2023). Reliable data enables businesses to make more informed decisions, improve marketing strategies, and increase customer engagement. Furthermore, the systematic sampling technique Combines churned and non-churned customers to improve model applicability across customer segments and maximize retention efforts.

However, the research has several flaws. The study concentrates solely on transaction frequency and value, ignoring demographics, purchase history, and engagement. This prevents businesses from identifying key customer segments for personalized marketing strategies, according to (Joung & Kim, 2023). It also does not consider important credit attributes: Credit scores, payment history, and risk profiles are not considered in churn prediction, which limits accuracy and targeted retention strategies.

While this study provides useful insights into churn prediction using machine learning, a more comprehensive approach that includes diverse customer data and long-term analysis would provide a more complete picture for developing effective customer retention strategies. Integrating credit-specific attributes would broaden the model's applicability in the credit card industry.

## 2.5. Literature 5:

Another literature conducted by (Bello, 2023) to investigate and evaluate how machine learning techniques can improve the accuracy, efficiency, and robustness of credit risk models. The primary goal was to determine how machine learning techniques can improve the precision, efficiency, and robustness of credit risk models when compared to traditional methods. The literature also seeks to address issues such as data privacy, model interpretability, and overfitting. Emerging trends like explainable AI (XAI) and federated learning are investigated for potential advantages.

The literature comprehensively evaluates various machine learning techniques, emphasizing their advantages over traditional models. It emphasizes novel approaches, such as real-time risk monitoring and alternative data sources. The emphasis on model interpretability using XAI techniques, as well as strategies for robust data privacy and security, is especially strong. In the finance industry model interpretability and transparency is very important, which is exactly what this study complies with and lays the foundation for future research to work with (Černevičienė & Kabašinskas, 2024). The forward-thinking exploration of future research

directions adds substantial value. In the context of credit card business optimization, these strengths contribute to the research question of improving model generalizability, and model interpretability using XAI techniques.

Despite its strengths, the literature contains significant flaws and gaps. There is a scarcity of real-world case studies and practical applications that demonstrate the effectiveness of machine learning models in financial settings (Nazareth and Reddy, 2023). While fairness and bias mitigation are discussed, empirical research on their real-world effectiveness is limited. These gaps highlight the need for additional research on the business impact analysis of various ML models. By comparing models to find the best balance of accuracy and recall, the potential impact of missed churners and defaulters can be assessed (Gupta et al., 2023). Furthermore, understanding the key factors influencing credit card customer churn and defaulting is critical for developing effective retention and risk management strategies.

In conclusion, this study on ML in credit risk assessment identifies significant advancements and future potential while also revealing significant gaps. To close these gaps, empirical research, real-world applications, and comprehensive regulatory frameworks are required. As a result, the financial industry can fully leverage machine learning to revolutionize credit risk assessment and create more robust, transparent, and equitable financial systems.

# Chapter 3: Methodology & Data

## 3.1. Literature Review Methodology:

A thematic literature review was utilised with a comparative approach to objectively examine the studies and identify strengths, weaknesses and gaps for each literature. A critical examination of the latest studies revolving around credit card customer churn and risk assessment methods is conducted. The scope of this research is broken down into two parts: technical and business aspects. Then, based on these two aspects, the studies were chosen using google scholar and ensuring that the selected research is of latest relevance. Lastly, the technological proposals and business implications of each study is analysed. The technological aspect focuses on the AI and ML algorithms used, along with relevant metrices to ensure good model and cluster performance. Whereas the business aspect is to ensure the interpretability, generalizability, and financial implications and strategies of the results to ensure these practises can be adopted in real world scenarios.

## 3.2. Approach To Analysis Overview:

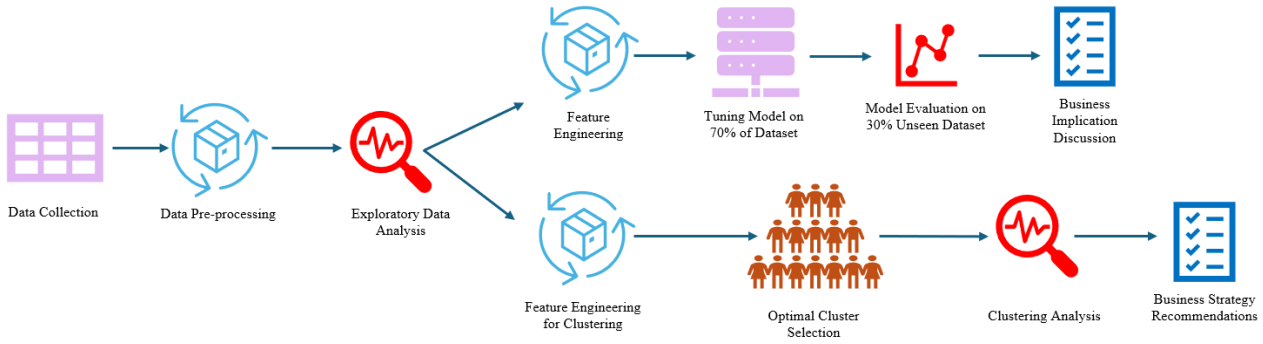


*Figure 1 - Overall Methodology Process*

This process is further subdivided to provide a deeper understanding of the modelling and clustering methodology.

### 3.2.1. Data Modelling Process:

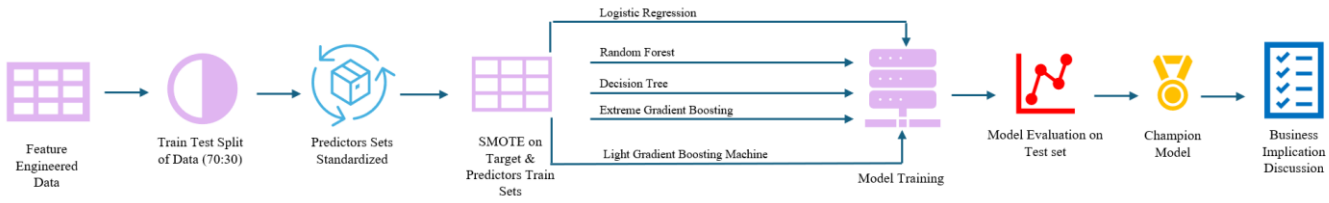


*Figure 2 - Data Modelling Process Breakdown*

### 3.2.2. Data Clustering Process:

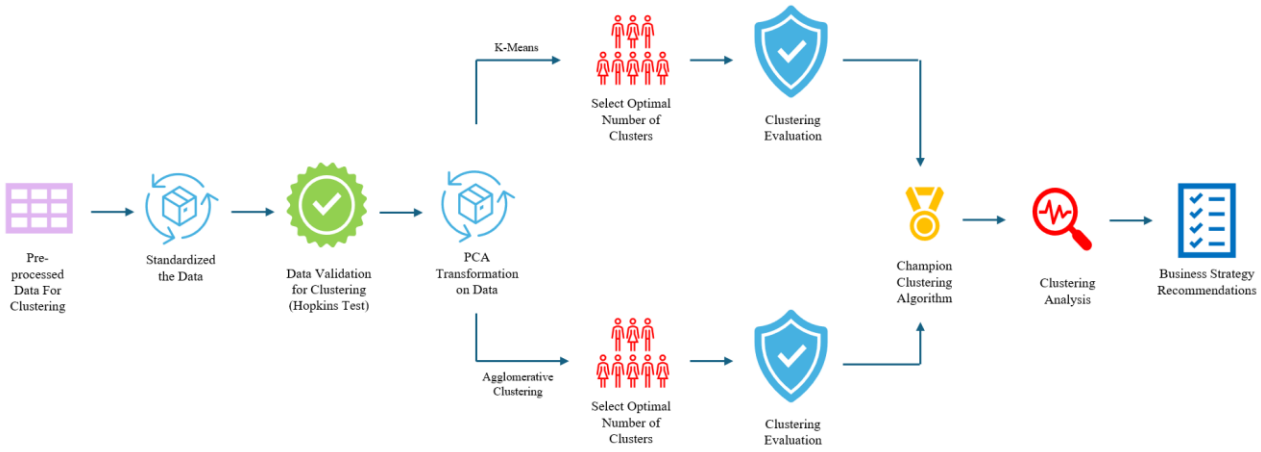


*Figure 3 - Data Clustering Process Breakdown*

Further on, this paper discusses the techniques used in the methodology for transforming, modelling, clustering and evaluating the data.

## 3.3. Feature Engineering Techniques:

### 3.3.1. Data Standardisation:

In machine learning, data standardization is a preprocessing step that converts the data's features to have a unit variance and zero mean (Santos, 2023). This procedure makes sure that every feature makes an equal contribution to the distance calculations, which is crucial for algorithms like Principal Component Analysis (PCA), Support Vector Machines (SVM), and Logistic Regression (LG) that are sensitive to feature scales.

Standardising data puts features in an equal scale, making the optimization or modelling process to be more efficient. It also helps prevent bias since some machine learning models are unable to capture the true meaning behind the features, if the scales are different. This in turn, leads to more accurate predictions and insights, which can improve the decision-making process for business strategies.

Please refer to Appendix 1 for technical details.

### 3.3.2. SMOTE (Synthetic Minority Oversampling Technique):

This method for achieving class distribution balance within a dataset. This approach concentrates on examples from the minority class, especially those that are close to the feature space boundary (Elreedy, Atiya, and Kamalov, 2024). This guarantees that the model can predict the minority class with accuracy and does not become biased in favour of the majority class. The SMOTE method is applied on the train set only, since applying it on test set would lead to leakage of data and overfitting (Imani and Arabnia, 2023). This helps the model to

generalize better on unseen data. Remarkably accurate minority class event identification and prediction can improve resource allocation and decision-making in business scenarios.

Please refer to Appendix 1 for the technical details.

### 3.3.3. PCA (Principal Component Analysis):

Principal Component Analysis (PCA) is a popular technique in data analysis for reducing the dimensionality of large datasets. Its primary goal is to reduce many variables to a smaller set that retains the most important information from the original data (Bruni, Cardinali, and Vitulano, 2022). This transformation is accomplished by identifying and creating new variables called principal components, which are linear combinations of the original variables. These components capture the dataset's maximum variance, condensing the information while minimizing information loss. PCA is used for clustering in this study. PCA enhances the efficiency and effectiveness of clustering algorithms by reducing the number of variables while preserving important patterns and trends (Bruni, Cardinali, and Vitulano, 2022). According to (Du, 2023), clustering algorithms frequently struggle with high-dimensional data due to increased computational complexity and the potential for dilution of meaningful patterns by noise. PCA addresses these issues by combining redundant data and increasing the signal-to-noise ratio, thereby improving the clustering process's accuracy and reliability.

Please refer to the Appendix 1 for technical details.

## 3.4. Machine Learning Algorithms:

### 3.4.1. Logistic Regression:

Logistic regression is a statistical model used to predict the likelihood of a binary outcome based on predictor variables. It's a generalized linear model (GLM) where the dependent variable is binary (0 or 1) (Zaidi and Luhayb, 2023). Logistic regression estimates the probability that an instance belongs to the positive class, offering clear interpretations of coefficients. These coefficients indicate the direction and strength of the relationship between each predictor and the outcome's log-odds, such as the likelihood of loan default or customer churn.

*The Formula:*

$$Log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

$Log\left(\frac{p(X)}{1-p(X)}\right)$ is the log odds of the probability *p(X)*.

$B_0$ is the intercept and $B_1, B_2$ etc. are the coefficients.

The logistic regression is one of the basic, yet effective machine learning classification algorithms. This study dives further into more advanced tree based and boosting algorithms that are used for modelling.

### 3.4.2. Decision Tree Classifier:

A decision tree is a supervised learning algorithm for classification and regression tasks. According to (Afriyie et al., 2023), it models decisions and their possible outcomes by breaking down complex decisions into simpler choices, making it highly interpretable. Decision trees are useful for credit risk assessment and customer churn prediction. They provide a visual representation of how different features lead to specific outcomes (Hada, Carreira-Perpiñán, and Arman Zharmagambetov, 2023). Leaf nodes indicate final predictions, branches indicate decision outcomes, and internal nodes represent feature-based decisions. Decision trees are scalable and can efficiently process large datasets, adapting to new data without extensive retraining. Decision trees make decisions by recursively splitting data using criteria like the Gini Index or Information Gain. The Gini Index measures impurity:

$$Gini(D) = 1 - \sum_{i=1}^{n} (p_i)\text{^}2$$

$p_i$ represents the probability of class $i$. Information gain measures the reduction in entropy:

$$IG(D, A) = Entropy(D) \quad - \sum_{u \in Values(A)} \frac{|D_u|}{|D|} Entropy\ (D_u)$$

### 3.4.3. Random Forest Classifier:

Random Forest is an ensemble learning algorithm for classification and regression tasks. It generates multiple decision trees during training and outputs either the mode of the classes or the average prediction. This ensemble approach enhances predictive accuracy and reduces overfitting, making it useful for credit risk assessment and customer churn prediction. According to the authors (Zang, Quost, and Masson, 2023), Random Forests provide higher accuracy than single decision trees, although their interpretability is more complex. Feature importance metrics help understand the underlying drivers of risk assessments or customer churn predictions (Özkurt 2024). Random Forests use bootstrap aggregation (bagging) to create multiple subsets of the training data by sampling with replacement (Sun et al., 2024). Each subset trains a separate decision tree, reducing variance and preventing overfitting. For classification tasks, each tree votes for a class, and the class with the most votes becomes the final prediction.

*The Formula:*

Random Forest using Gini Index:

$$Gini = 1 - \sum_{i=1}^{C} (p_i)\text{^}2$$

This formula calculates the Gini of each branch on a node using the class and probability to identify which branch has a higher probability of occurring. In this case, $c$ is the number of classes, and $p_i$ is the relative frequency of the class you are observing in the dataset.

Random Forest using Entropy:

$$Entropy = \sum_{i=1}^{C} - p_i * log_2(p_i)$$

Entropy decides how the node should branch based on the likelihood of a particular result. Because of the logarithmic function that is utilized to calculate it, it is more mathematically complex than the Gini index.

### 3.4.4. Diagrammatic Example of Bagging Classification (Random Forest)

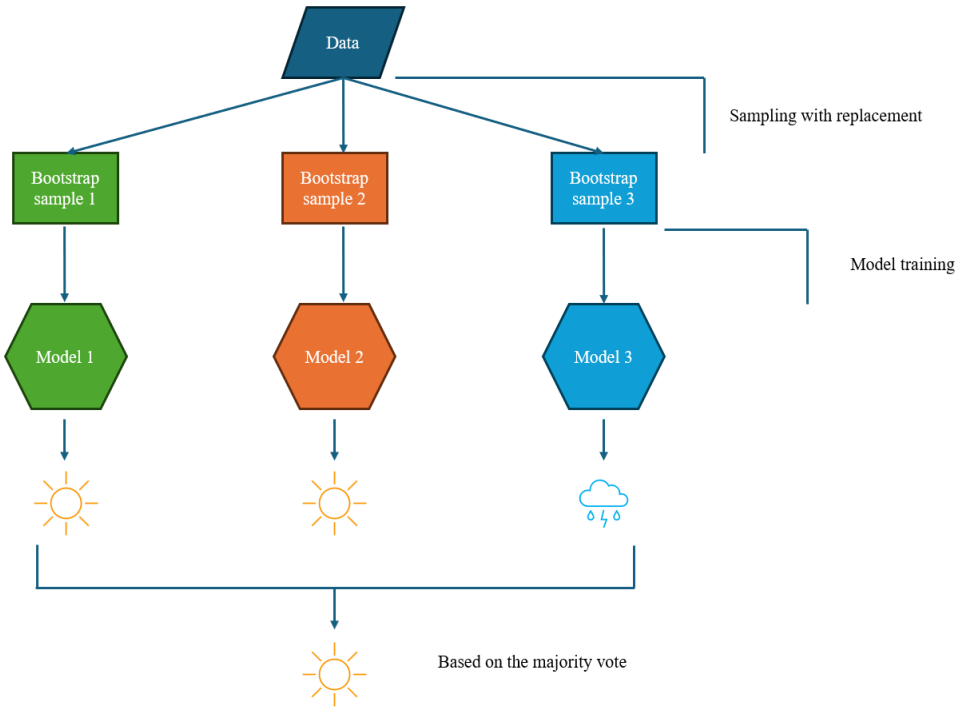


*Figure 4 - Random Forest Process*

### 3.4.5. Extreme Gradient Boosting Classifier:

Extreme Gradient Boosting (XGBoost) is a highly efficient and accurate machine learning algorithm. Unlike simpler models, XGBoost uses an ensemble of decision trees, sequentially correcting errors to enhance predictive accuracy (Ali et al., 2023). This makes it ideal for precise predictions in banking and credit card applications. XGBoost optimizes an objective function with a loss function to compute prediction errors and regularization to prevent overfitting (Tarwidi et al., 2023). While its ensemble nature may complicate interpretability, feature importance analysis can highlight key factors like income or credit history. XGBoost's scalability and ability to process real-time data make it well-suited for dynamic business environments, allowing it to adapt to new information without extensive retraining.

The objective loss function $L$ at iteration $t$ that needs to be minimized:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

### 3.4.6. Light Gradient Boosting Classifier:

LightGBM (Light Gradient Boosting Machine) is a cutting-edge gradient boosting framework known for its speed, effectiveness, and reliable performance in machine learning applications (Omotehinwa, Oyewola, and Dada, 2023). It introduces innovative features like Gradient-based One-Side Sampling (GOSS) to focus on data instances with large gradients for better precision, and Exclusive Feature Bundling (EFB) to optimize feature handling and reduce memory usage. LightGBM's scalability allows processing millions of data points without a proportional increase in resources. According to (Hajihosseinlou, Maghsoudi, and Ghezelbash, 2023), unlike XGBoost, LightGBM uses a histogram-based approach for decision tree construction, improving speed and memory efficiency. It also employs a leaf-wise growth strategy, building deeper trees for better performance. LightGBM's hierarchical tree structures enhance interpretability, making decision processes transparent and aiding regulatory compliance.

### 3.4.7. Diagrammatic Example of Boosting Classification (Boosting Algorithms):

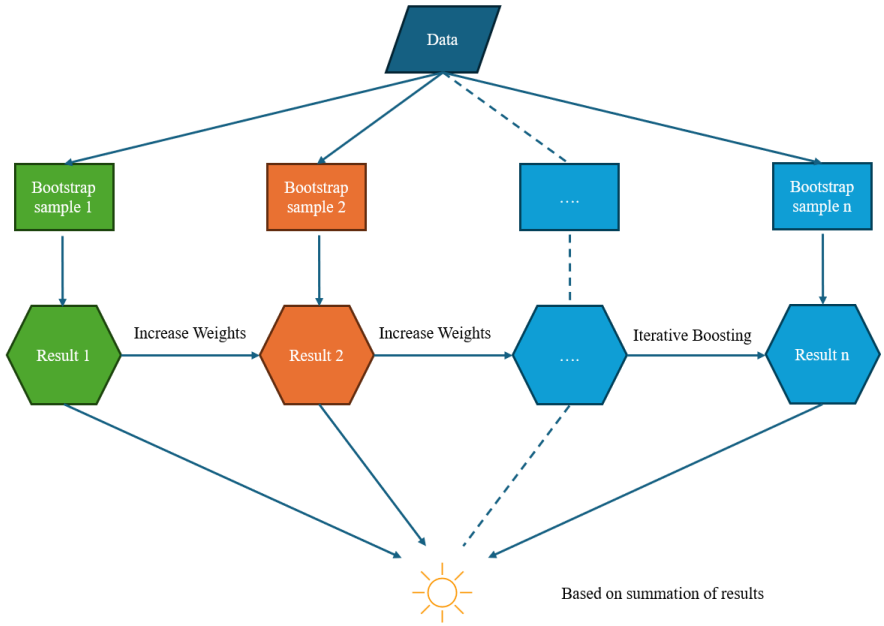


*Figure 5 - Boosting Algorithms Process*

## 3.5. Clustering Algorithms:

### 3.5.1 K-means Clustering:

K-means clustering is an unsupervised machine learning algorithm that partitions a dataset into K distinct clusters (John, Shobayo, and Ogunleye, 2023), Initially, *K* centroids are placed randomly or using a heuristic method. In each iteration, data points are assigned to the nearest centroid based on distance metrics like Euclidean distance. Centroids are recalculated as the mean of the assigned data points until convergence, aiming to minimize variance within clusters. K-means has a time complexity of *O(n * K * I)*, where n is data points, *K* is clusters, and *I* is iterations.

This research employs the use of clustering to reveal hidden patterns regarding customer behaviour. However, choosing the optimal K requires careful parameter tuning and initialization (Ikotun et al., 2023). K-means can be quite sensitive to initial placement of centroids and can struggle with very large datasets. Fortunately, the size of the data used in this study are manageable.

Diagrammatic Example:

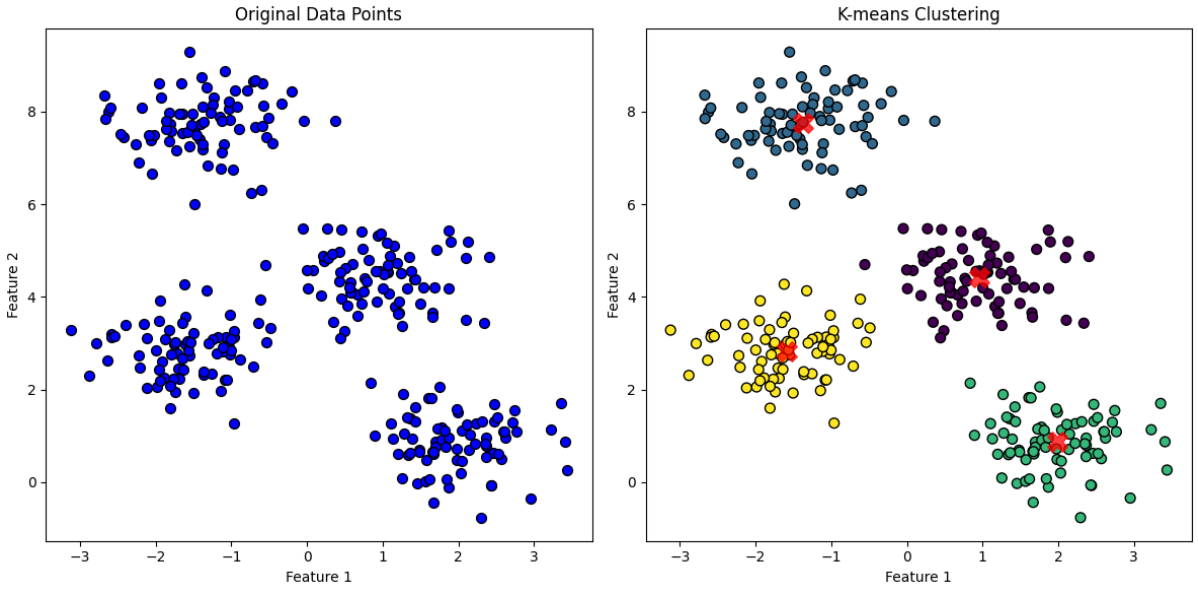


*Figure 6 – K-means Clustering Representation*

### 3.5.2. Hierarchical Clustering:

Agglomerative hierarchical clustering groups data into a hierarchy of clusters by iteratively merging the closest clusters. According to (John, Shobayo, and Ogunleye, 2023), initially, each data point is treated as its own cluster. The algorithm computes pairwise distances between clusters using metrics like Euclidean distance or correlation, merging the closest clusters until a specified number of clusters or a distance threshold is met. This process results in a dendrogram, a tree-like structure that illustrates the merging order and distances (John, Shobayo, and Ogunleye, 2023), In this research, hierarchical clustering is valuable for

revealing hierarchical relationships within data, aiding market segmentation by identifying customer groups with similar characteristics. It supports customized marketing strategies and resource allocation by organizing departments or processes based on similarities. Despite its computational complexity, hierarchical clustering offers interpretable insights into relationships between variables or products, enhancing strategic decision-making and planning.

Diagrammatic Example:

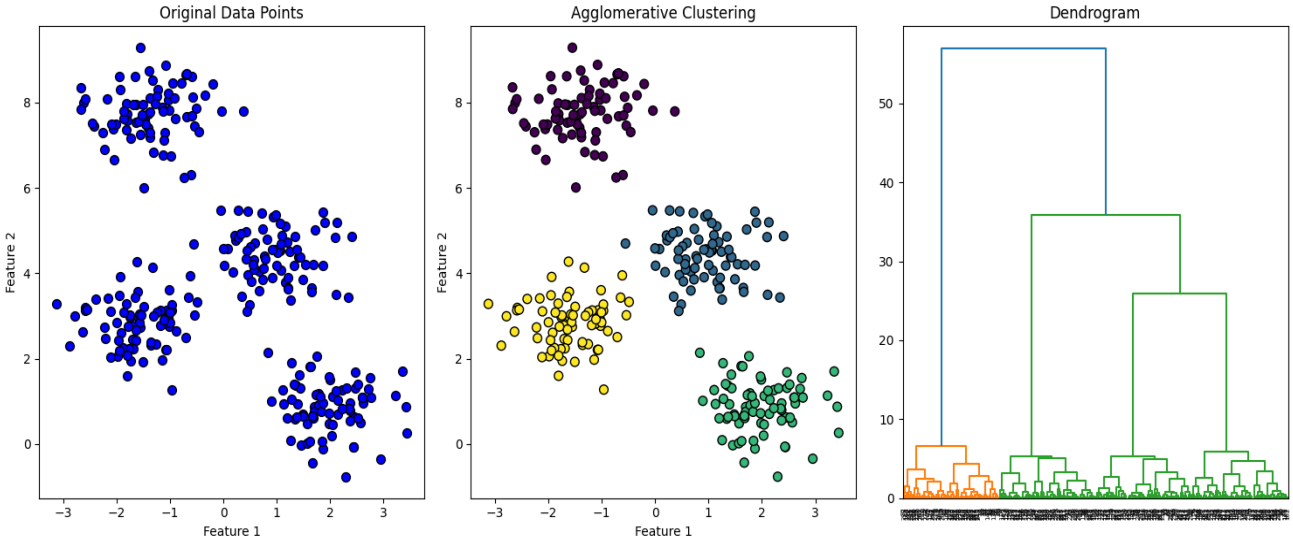


*Figure 7 - Agglomerative Clustering Representation*

## 3.6. Evaluation metrices:

### 3.6.1. Precision:

Precision is the proportion of correctly predicted positive instances (true positives) among all instances predicted as positive. It's computed as:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

A high precision means that the model is probably right when it predicts a positive outcome.

### 3.6.2. Accuracy:

The percentage of correctly predicted instances—both true positives and true negatives—out of all evaluated instances is known as accuracy. It is computed as follows:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Instances}$$

It is an overall indicator of the model's performance across all classes.

### 3.6.3. Recall:

Recall calculates the proportion of correctly predicted positive instances (true positives) among all actual positive instances. It's computed as:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

A high recall rate means that most positive instances can be identified by the model.

### 3.6.4. F1:

The F1 score is a metric that provides a balance between precision and recall, calculated as the harmonic mean of both measures. The calculation is as follows:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The F1 score is useful for keeping a balance between precision and recall, particularly when classes are imbalanced.

### 3.6.5. Davies Boulding Index:

The Davies-Bouldin Index calculates the average dissimilarity between clusters in relation to the average similarity between each cluster and its most similar cluster. The calculation is as follows:

$$DBI = \frac{1}{K}\sum_{i=1}^{K} max_{j!=i}\left(\frac{(\sigma_i + \sigma_j)}{d(c_i, c_j)}\right)$$

Here $K$ is the number of clusters. $c_i$ and $c_j$ are the centroids of the $i$ and $j$ clusters. $\sigma_i$ and $\sigma_j$ are the average distances from each point in clusters $i$ and $j$ to their centroids, and $d(c_i,c_j)$ is the distance between both the centroids $c_i$ and $c_j$.

### 3.6.6. Silhouette Score:

The degree to which a point resembles its own cluster in relation to other clusters is determined by its Silhouette Score. It is computed as follows for every sample $i$:

$$Silhouette(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The average distance between a point and other points in the same cluster is denoted by *a(i),* while the average distance between a point and points in the closest neighbouring cluster is represented by *b(i).*

### 3.6.7. Calinski-Harabasz Index:

The Variance Ratio Criterion, or Calinski-Harabasz Index, measures the ratio of the total between-cluster dispersion to the within-cluster dispersion. It is computed as follows:

$$CHI = \frac{Tr(B_k)}{Tr(W_k)} * \frac{N - K}{K - 1}$$

*Tr* represents the trace of the between-cluster scatter matrix. *Bk* is the between-cluster scatter matrix, *Wk* is the within-cluster scatter matrix, with *N* being the total number of samples and *K* is the number of clusters.

## 3.8. Data Description:

There are two datasets used for this study primarily.

The first is the credit card customers dataset from Kaggle by (Goyal, 2020). This dataset comprises of 10,127 customer records with 21 features. It has 0 duplicates and no null values. The dataset is first feature engineered for clustering purposes. Only the numerical variables in the dataset are selected for clustering, as numerical data is most optimal for the clustering process and reveals hidden insights on customer behaviour behind the numbers. Next, the dataset is pre-processed for modelling. The transformed dataset has 38 variables and 10,127 records of customer data.

The second dataset used is the Taiwan credit card default dataset from the UCI repository by (Yeh, 2016). This dataset consists of 30,000 records of customer data with 25 variables in total. It has 35 duplicates and no null values and is utilized as the evaluation data for predicting credit card default. This dataset extracts new variables: the credit utilization rate for each month based on the 'LIMIT_BAL' (credit limit) and the 'BILL_AMTX' (monthly balance) variables. The average credit utilization rate is also extracted from the credit utilization rate of each month. A new variable is created that flags customers that have above 100% average credit utilization rate. The total payment amount variable is calculated by adding all the 'PAY_0' variables. The

total payment relative to credit limit variables is created from the "total payment amount" and 'LIMIT_BAL' variables. Lastly, a new variable is created that flags customers with overpayments. The data transformations lead to total of 117 duplicates which are removed to reduce bias during modelling process. The dataset is then feature engineered for modelling and clustering and has 29,823 customer records with 43 variables.

The currency used for financial context in this study is NT$ (New Taiwan Dollars), to ensure simplicity and consistency among both datasets.

Please refer to Appendix 2 for information on variables in the dataset and detailed explanation of data transformation steps.

### 3.8.1. Class Imbalance:

This section presents the class imbalance of the target variable for both the datasets.

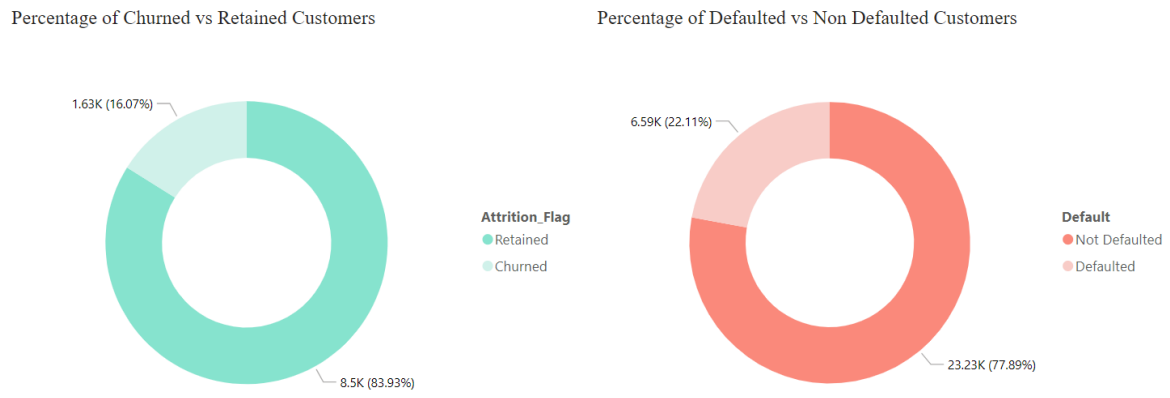


*Figure 8 - Class Imbalance Distribution in Both Datasets*

From the above graph it can be observed that in the churn dataset, the percentage of churned customers is less than retained by 67.86%, similarly for the default dataset, the percentage of defaulted customers is less than non-defaulted customers by 55.78%. This class imbalance is addressed using SMOTE in the analysis.

# Chapter 4: Analysis & Insights

This chapter aims to provide information based on the insights gathered from data clustering and modelling for both datasets. First the optimal cluster selection is reviewed using Hopkins value for assessing the clustering tendency of a dataset, after which the optimal clusters are selected and described. Furthermore, exploratory analysis is conducted with a series of graphs, which are followed up with detailed summaries of the graphs altogether. Finally, the ML model results are compared for both datasets separately.

## 4.1. Clustering Analysis:

This section focuses on selecting optimal number of clusters using K-means and agglomerative clustering for both datasets separately.

### 4.1.1. Credit Card Customer Churn:

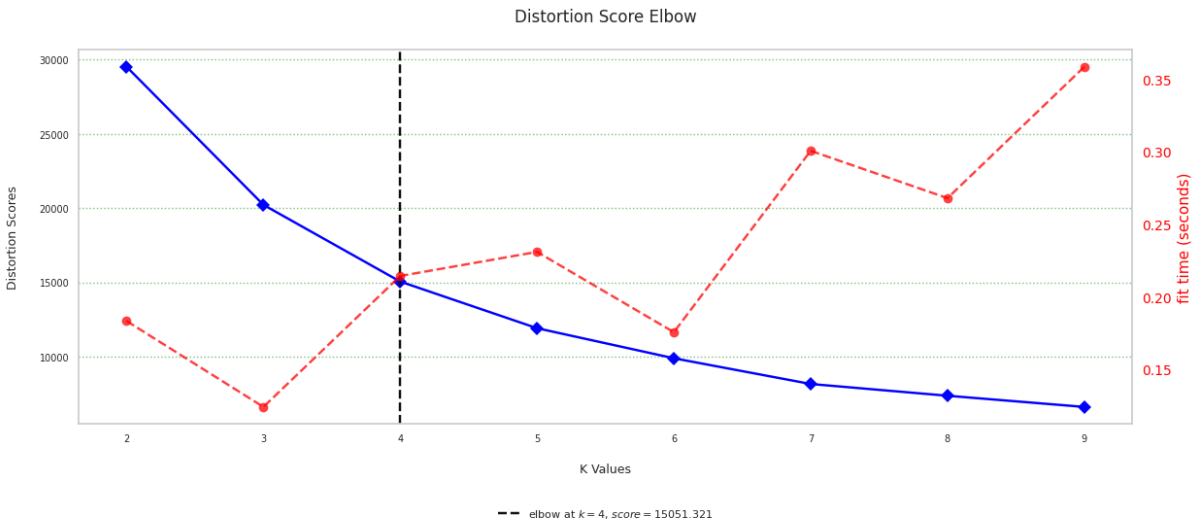


*Figure 9 – Optimal Clusters for K-means*

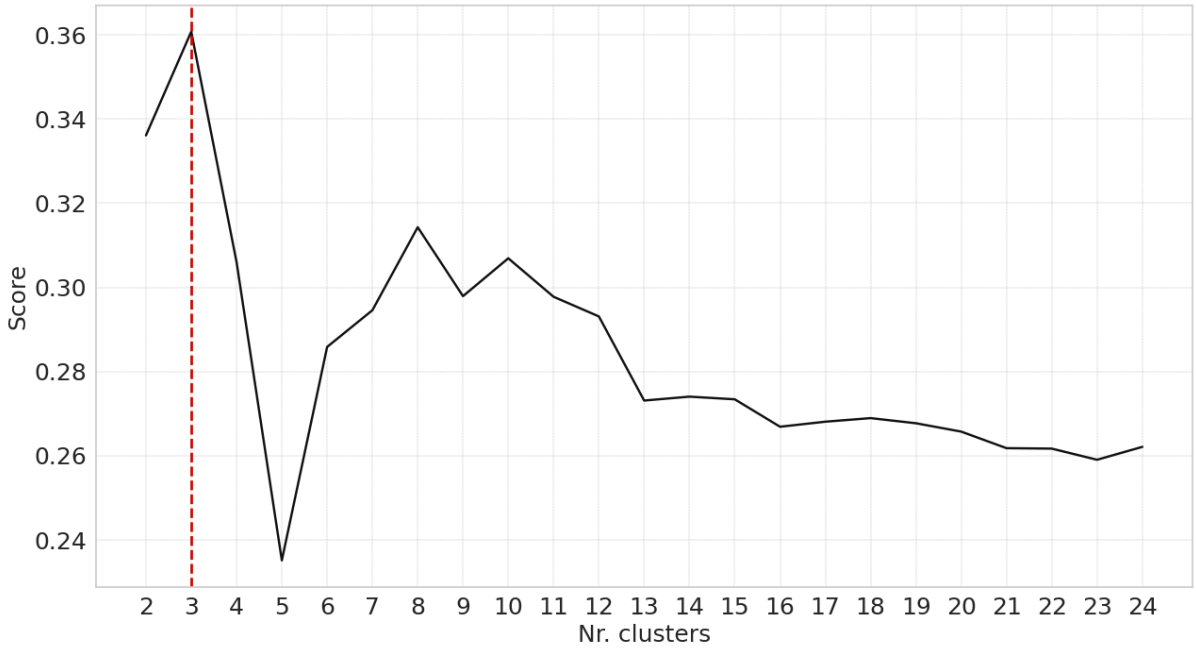


*Figure 10 - Optimal Clusters for Agglomerative Clustering*

After extracting and transforming numerical features using standardisation and PCA, the dataset had a Hopkins value of 0.802, indicating a strong tendency to form clusters suitable for customer segmentation. A comparison of the optimal number of clusters between two algorithms revealed that K-means suggested four clusters, whereas agglomerative hierarchical clustering suggested three.

| Evaluation Metrices | K-Means | Agglomerative Clustering |
|---|---|---|
| Davies Boulding Index | 0.9 | 1.002 |
| Silhouette Score | 0.373 | 0.361 |
| Calinski-Harabasz Index | 7099.29 | 5690.596 |

*Table 1 - Clustering Techniques Comparison for Churn*

The K-means method, with a higher Silhouette Score (0.373), lower Davies-Bouldin Index (0.9), and higher Calinski-Harabasz Index (7099.29), was chosen as the best method, indicating the selection of four clusters. Hence the optimal number of clusters to select is 4.

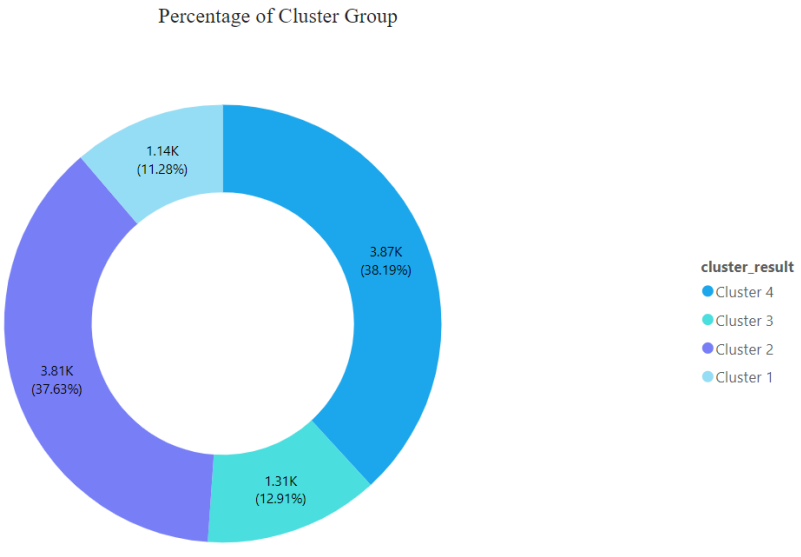


*Figure 11 - Percent Distribution of Clusters (Churn)*

The above graph showcases each cluster count and its percentage. Cluster 4 is the largest, consisting of around 3,870 customers, that makes up 38.19% of the total customer population. Second place is cluster 2 in terms of number, comprising of approximately 3,810 customers. Clusters 1 and 3 are relatively much smaller than 2 and 4. Both consisting of less than 15% of total customers, with cluster 1 being the smallest.

Each cluster has been briefly described below:

**Cluster 1 (Budget-Conscious Families):** Includes young families (average age 44, 2-3 dependents on average). Moderate engagement, 15% credit utilization, and consistent transaction behaviour. The overall transaction amount is high (NT$12k).

**Cluster 2 (Loyal Long-Term Customers):** Oldest customers (highest average age of 49, few dependents). Longest tenure, highest relationship count (4 products). Moderate engagement and low credit utilization (14%). Less transaction activity (NT$2.8k).

**Cluster 3 (High-Net-Worth Customers):** Has the most dependents (2 but mostly 3 on average), as well as the second highest average age (47). Long tenure with high engagement. Highest credit limit (NT$26k), with very low credit utilization (3%). Moderate transaction volume (NT$3.8k).

**Cluster 4 (Credit-Reliant Rising Spenders):** Includes young families (average age 44, 2 dependents on average). Moderate engagement. Credit utilization was the highest (52%), and transaction volume increased the most (NT$3.9k). The lowest credit limit (NT$3.4k).

For descriptive statistics regarding each cluster, please see Appendix 3.

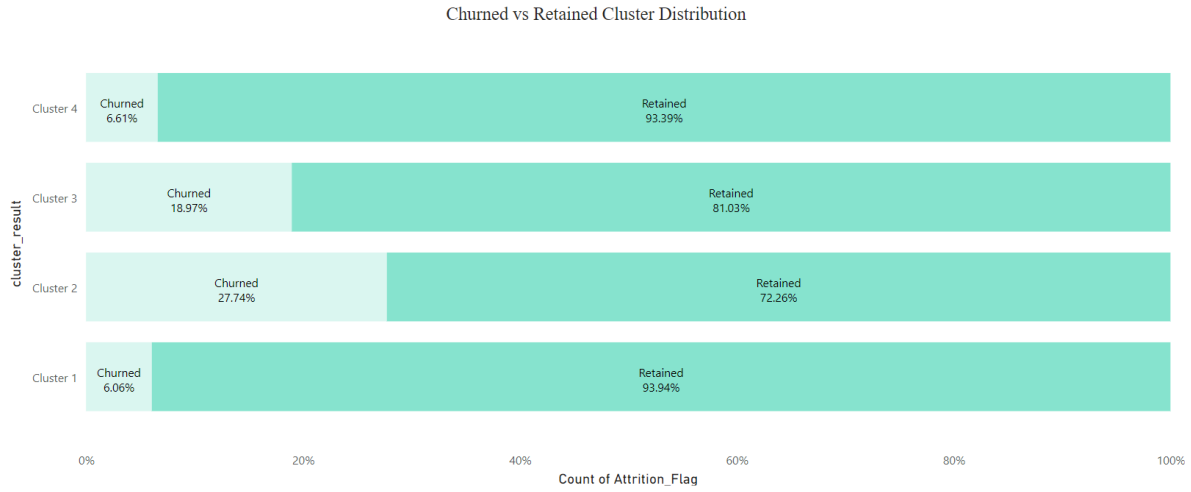


*Figure 12 - Attrition Distribution in each Cluster*

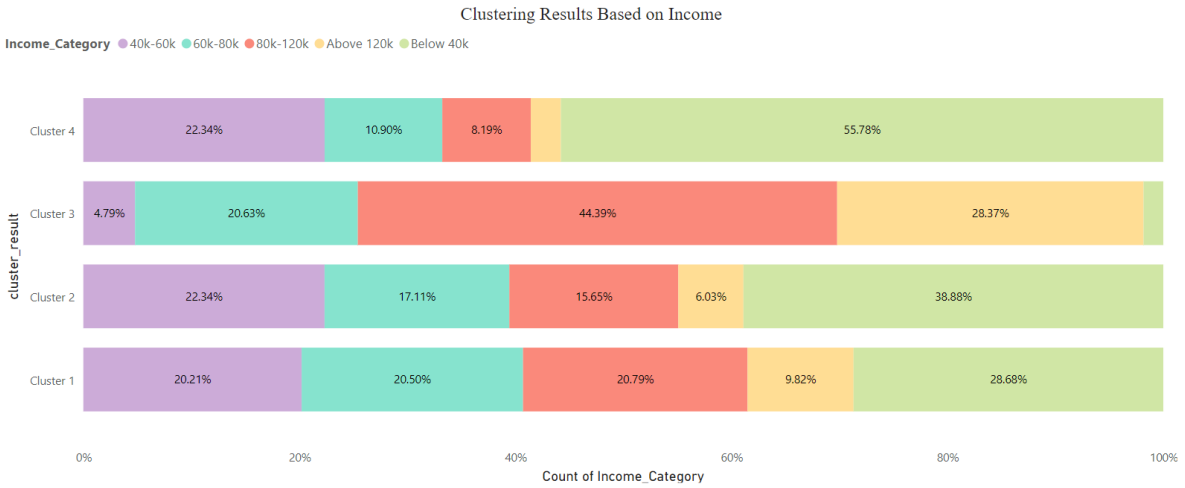


*Figure 13 - Income Category Distribution in each Cluster*

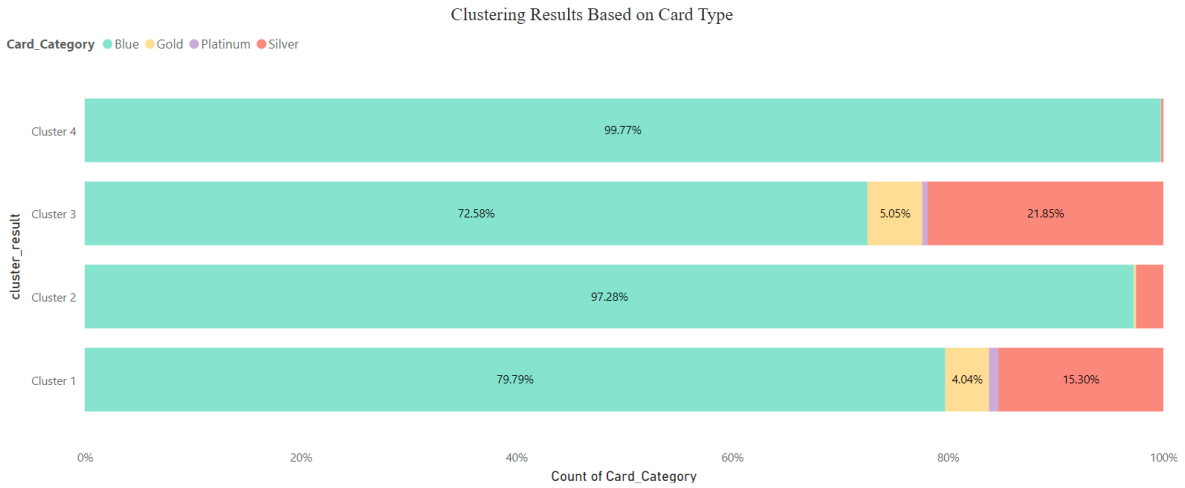


*Figure 14 - Card Type Distribution in each Cluster*

Based on the above three stacked graphs:

**Cluster 1:** Has the highest retention rate, which indicates high customer satisfaction. They have been with the bank for an average of 34 months and come from young families (average age 44) with moderate family responsibilities. Despite moderate credit limits and balances, they have the highest overall transaction activity and a preference for premium cards (80% blue, 0.9% platinum). Their balanced income distribution (20% each in the 40-60k, 60-80k, and 80-120k income brackets, with 10% above 120k) indicates a stable and diverse customer base with a strong preference for the credit card company offerings.

**Cluster 2:** Has the highest churn rate (27.92%), indicating a potential need for improved engagement strategies, particularly among long-term customers, who have the oldest average age and tenure. Their low credit limits, moderate transaction activity, and significant proportion of low-income earners (38.88% earning less than NT$40,000) indicate financial constraints and limited access to premium products (97.61% blue cards).

**Cluster 3:** Despite having a higher retention rate than Cluster 2, this segment loses 8.67% of customers. Financially stable (high credit limits, low utilization), these customers have a wide income distribution (nearly 44% earn 80k-120k, with 28.37% earning more than 120k). Their preference for premium cards (5% gold, 22% silver) corresponds to their higher income. Despite having lower overall activity than Cluster 1, their financial profile indicates that they may be at risk of churn and should focus on retention.

**Cluster 4:** Consists of the second-highest retention rate, but the highest proportion of low-income customers (more than half earn less than NT$40k). This suggests financial constraints. The lowest average credit limit, combined with the highest credit utilization ratio, demonstrates a greater reliance on available credit. Despite these challenges, they engage moderately in transaction activity and bank relationships. Their reliance on blue cards (99.79% of population) demonstrates a focus on basic banking services. This segment presents an opportunity to broaden product offerings to better meet their specific requirements.

### 4.1.2. Credit Card Default:

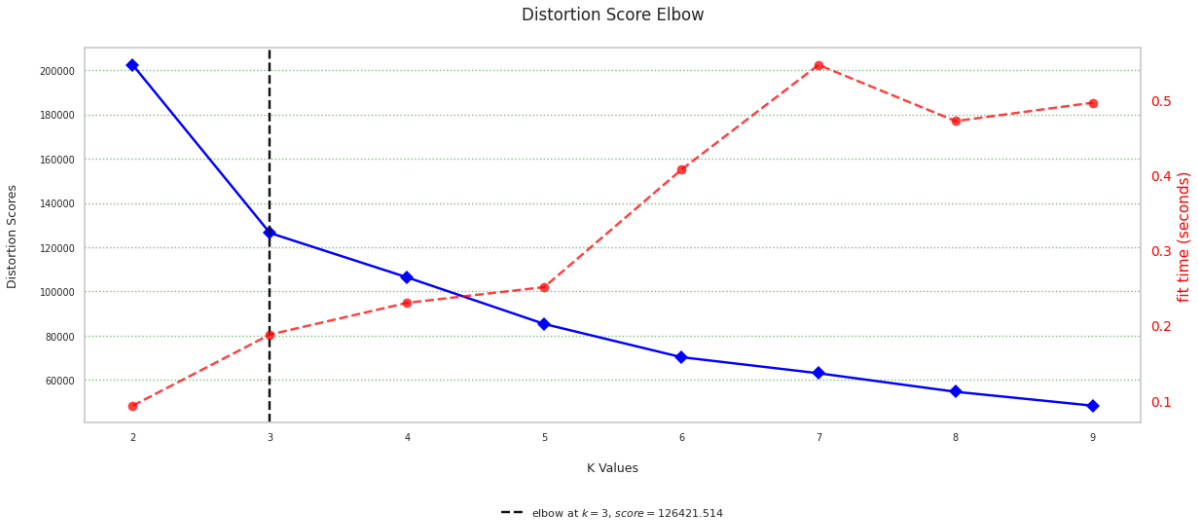


*Figure 15 - Optimal Clusters for K-means*

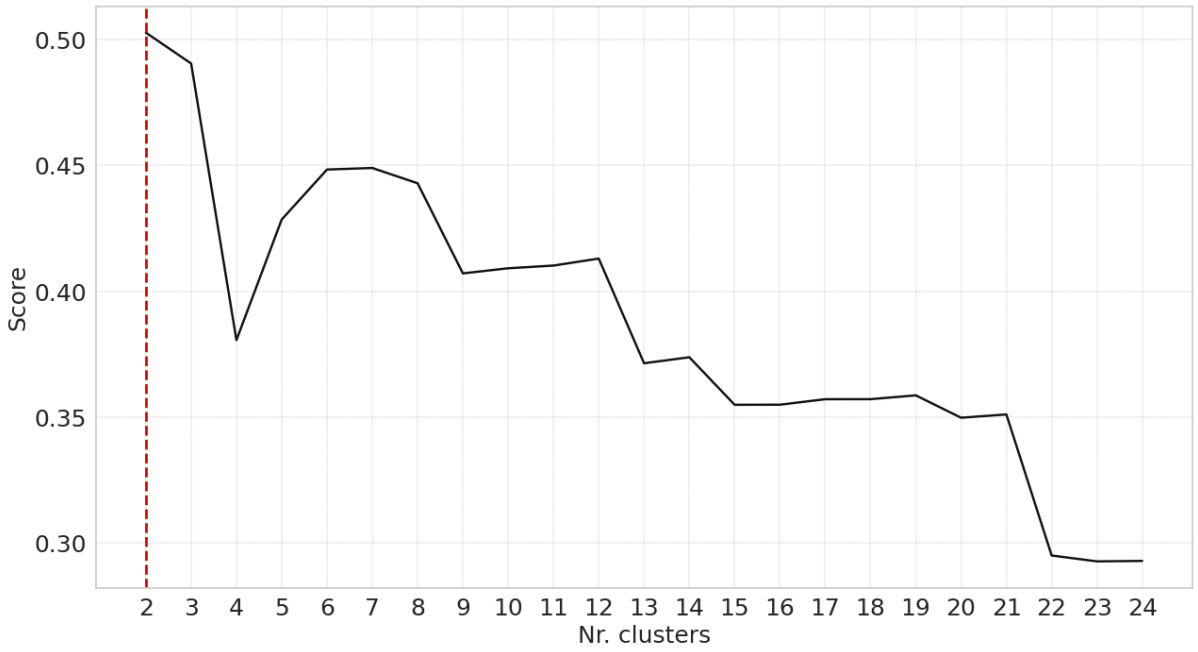


*Figure 16 - Optimal Clusters for Agglomerative Clustering*

The credit card default dataset received a Hopkins of 0.99, after only the numerical features were extracted and transformed using standardisation and PCA. This value is extremely close to 1, thus implying that the data has very strong tendency to form clusters. As observed from the above graph, the elbow method which is the K-means method, suggests 3 clusters. In contrast the agglomerative hierarchical clustering method suggests 2 clusters. Upon testing both methods using the clustering evaluation metrices, these are the following results:

| Evaluation Metrices | K-Means | Agglomerative Clustering |
|---|---|---|
| **Davies Boulding Index** | 0.761 | 0.841 |
| **Silhouette Score** | 0.546 | 0.503 |
| **Calinski-Harabasz Index** | 32117.727 | 26049.41 |

*Table 2 - Clustering Techniques Comparison for Default*

The table compares K-Means and Agglomerative Clustering with three evaluation metrics. K-Means outperforms Agglomerative Clustering by having a lower Davies-Bouldin Index (0.761 vs. 0.841), a higher Silhouette Score (0.546 vs. 0.503), and a higher Calinski-Harabasz Index (32177.727 vs. 26549.51). These findings show that K-Means produces more distinct, well-defined clusters with better dispersion between clusters, indicating overall superior clustering performance. Thus, the optimal number of clusters selected is 3.

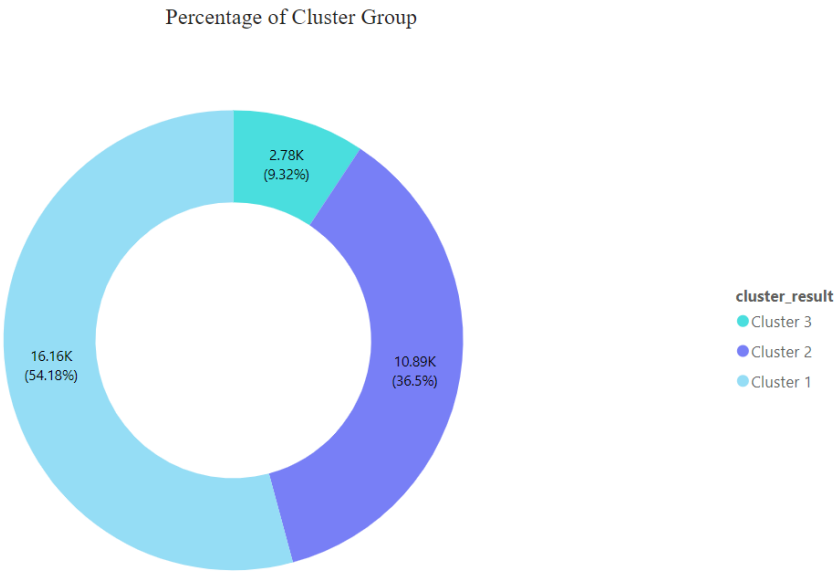


*Figure 17 - Percent Distribution of Clusters (Default)*

The above donut chart showcases the population segments of the clusters in percentage. Cluster 1 makes up more than 50% of the total customer data consisting of approximately 16160 customers. Cluster 2 is the second largest, making up 36.5% of the total customer population, with the smallest being cluster 3 (9.32%).

Each cluster has been briefly described below:

**Cluster 1 (Financially Savvy Spenders):** Consistent payments of approximately NT$4,500 and high credit limits (NT$200k+) with low utilization (9%) indicate financially responsible clients.

**Cluster 2 (Financially Strained Payers):** Lower payments (about NT$3,200) and high utilization (71%) of moderate credit limits (NT$80k) point to possible financial strain.

**Cluster 3 (Risk-Taking Credit Consumers):** The highest credit limits (NT$300k+) are correlated with high payments (approximately NT$19,000) and heavy credit use (71%)—high revenue potential but also higher risk.

For descriptive statistics regarding each cluster, please refer to Appendix 3

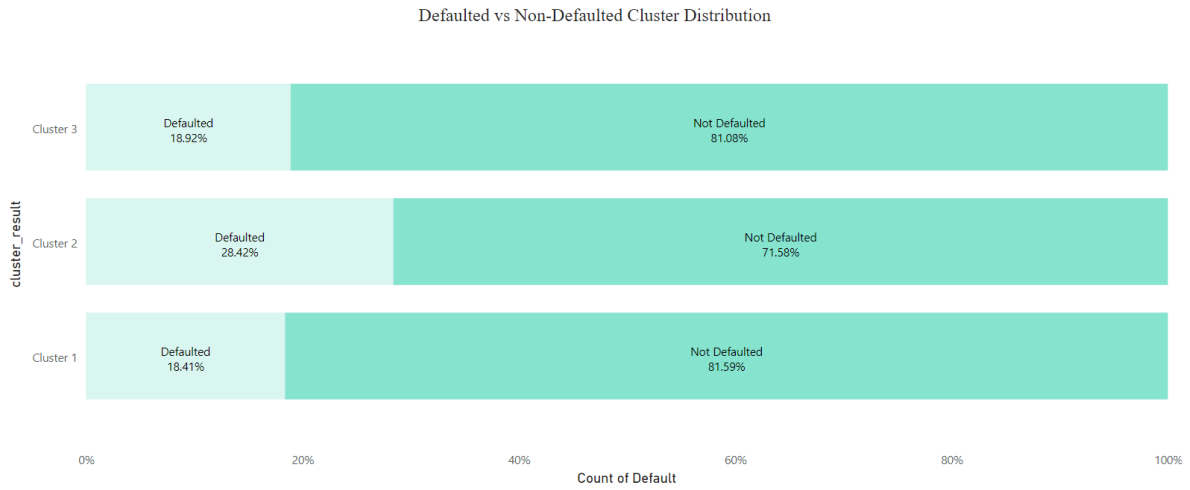


*Figure 18 - Default Distribution in each Cluster*

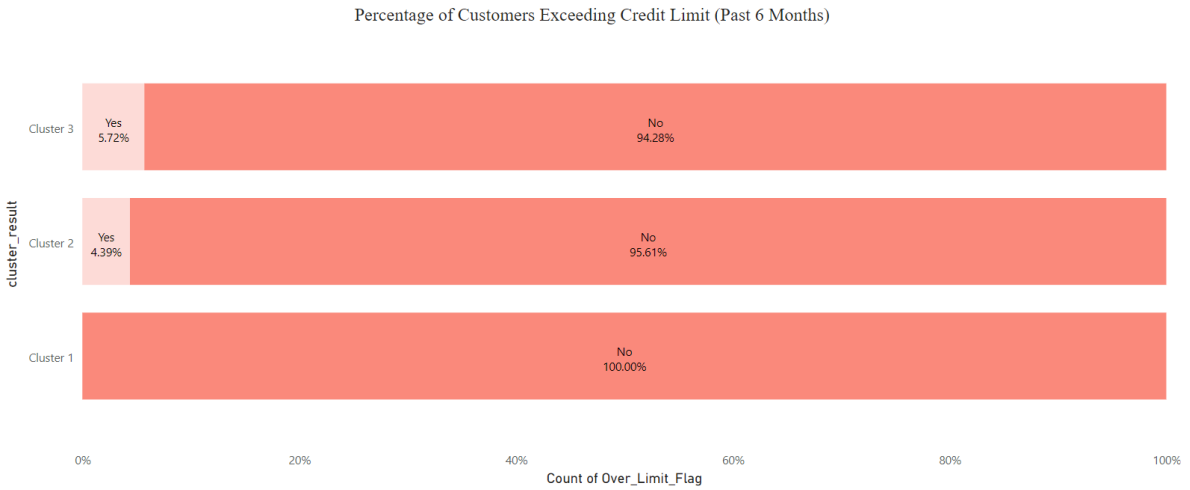


*Figure 19 - Distribution of Customers Exceeding their Credit Limit in each Cluster*

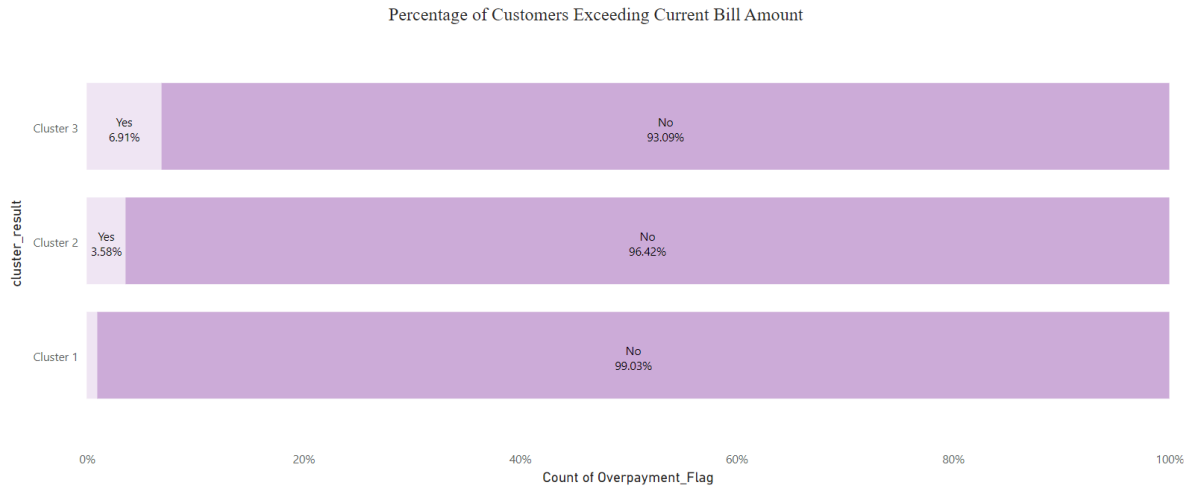


*Figure 20 - Distribution of Customers Exceeding Current Bill Amount in each Cluster*

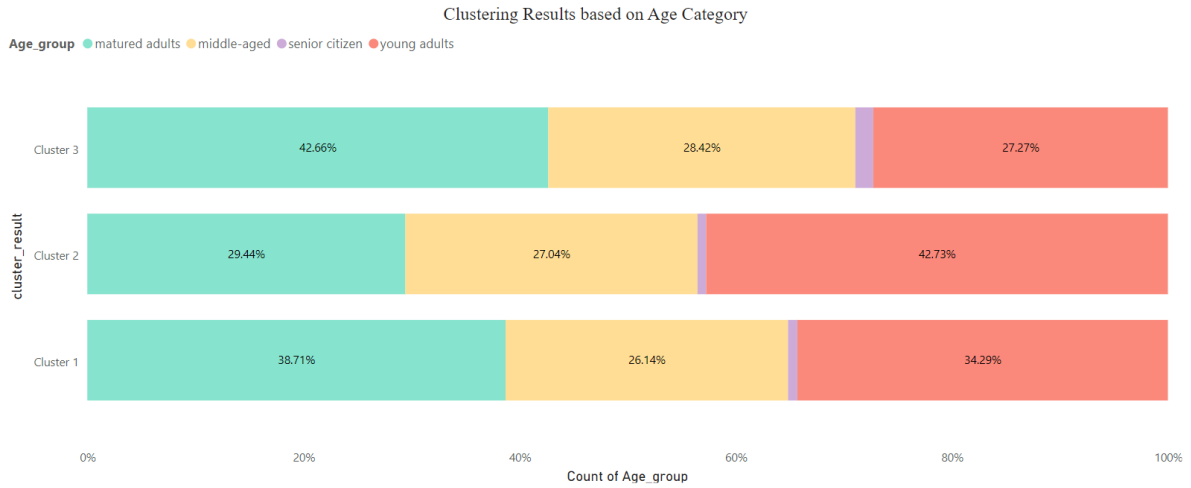


*Figure 21 - Age Category Distribution in each Cluster*

From the above four stacked graphs, the following insights are extracted:

**Cluster 1**: Has a more balanced and financially responsible profile, with a default rate of 18.41% and an impressive non-defaulting percentage of 81.59%. Notably, no customers in Cluster 1 have exceeded their credit limits, and less than 1% have exceeded their bill amounts, indicating effective credit management. This cluster is made up of 39% mature adults (30-40 years), 34% young adults (20-30 years), and 26% middle-aged adults (40-60 years). Senior citizens (60+) account for 0.86% of this group. The demographic distribution indicates a customer base that is relatively mature and financially stable.

**Cluster 2**: Emerges as the riskiest group, with a default rate of 28.42%, significantly higher than the other clusters. These customers are more likely to default, which is consistent with their financial behaviour. Cluster 2 has a high rate of financial overextension, with 4.39% of customers exceeding their credit limits and approximately 4% exceeding their current bill balances. This cluster has the highest proportion of young adults (42.73%), who may be less experienced in credit management, contributing to higher default rates. This cluster is made up of 30% mature adults (30-40 years) and 27% middle-aged adults (40-60 years). The senior citizen representation is minimal, at 0.8%.

**Cluster 3**: Has similar default rates to Cluster 1, with 18.92% defaulting and approximately 81% non-defaulting. However, Cluster 3 has slightly riskier behaviours, with approximately 5.7% exceeding their credit limits and 7% exceeding their bill amounts, the highest of any cluster. This cluster has the highest proportion of mature adults (30-40 years) at 43%, implying that financial obligations and expenses may be higher in this age range. 27.2% of the population is under the age of 30, while 28.5% is between the ages of 40 and 60. Senior citizens account for only 1.5%.

## 4.2. Modelling Analysis & Comparison:

Due to the high imbalance classification of target variables present in both datasets, this study directly presents the results using SMOTE, thus addressing class imbalance.

The results for the non-SMOTE models are given in Appendix.

### 4.1.1. Credit Card Customer Churn:


*Figure 22 - Churn Model Comparison*

**Model Performance Comparison:**

The above graph showcases the performance of the fine-tuned models for predicting churn after testing.

Logistic Regression: Shows high recall, precision and F1 in general for the negative class (Retainers). While that is valuable for necessary resource allocation, missing churners are of higher priority. It is unable to get higher scores for the positive class (Churned) in all areas, which can cause potential financial losses in real life scenario. It has a good accuracy of 90% although still the lowest of all the other models.

Random Forest: This model shows far better results. Its recall for churners is 85% with high recall for retainers as well. Its precision for churners and retainers are also quite high (84% and 97%), which means it is accurately able to identify false positives and negatives. With a high F1 result for both classes, and an accuracy of around 95%, this model is robust and ensures good generalizability on unseen data.

Decision Trees: Provides higher recall on churners than the random forest, although it is inferior to the random forest model in all other areas. While it is still overall a decent model, it could improve its precision for identifying churners (77%). However, its high accuracy (93.5%) ensures the model's robustness and generalizability.

Extreme Gradient Boosting: Gradient Boosting machines such as the XGBoost seem to generalize well in terms of churning data. It provides high accuracy, precision, recall and f1-score in for identification of both churners and retainers. Despite its lower recall for churners (0.84), it is still higher than the other models in every aspect except the LightGBM.

Light Gradient Boosting: This Boosting model provides the best result out of all the models. With an accuracy of around 96%, it is successfully able to identify non churners and churners effectively with a score of 85% + in all metrices for identifying the positive class accurately.

### 4.2.2. Credit Card Default:


*Figure 23 - Default Model Comparison*

**Model Performance Comparison:**

The above graph showcases the performance of the fine-tuned models for predicting default after testing.

Logistic Regression: This fine-tuned model results in high accuracy of around 80%, but it lacks in crucial metrices such as low recall (0.42) for defaulters. This can cause significant financial loss due to the model being unable to detect those who miss payments monthly. The model excels in identifying non defaulters. However, classifying a defaulter as non-defaulter is far more costly, than vice versa.

Random Forest: The model offers a stronger balance with Precision (0.51) and Recall (0.58) for identifying defaulters. This model effectively balances the detection of defaulters while maintaining precision, ensuring fewer false positives and minimizing risk. Although it does not have the best accuracy (78%), it still provides the best trade-off between recall and precision which makes it a robust model.

Decision Trees: In this, the accuracy is quite high at around 80%, beating the random forest model. However, the model isn't robust enough to identify defaulters as its recall is just 0.47. The results are bias towards detecting non-defaulters accurately, which is good, but identifying defaulters is more important.

XGBoost: Interestingly, the XGBoost model in this case has the least robustness out of all the models. It provides the lowest accuracy of 76%, and the lowest precision, recall and F1 for identifying the defaulters. It has high tendency to misclassify the defaulters as non-defaulters, due to its bias in results towards the negative class.

LightGBM: This model has the best accuracy (0.81) compared to the other models. It Its precision, for churners is approximately 61%, thus correctly predicting the positive class over 60% of the time. In terms of precision-recall balance, the f1 score is 50% for defaulters. However, for just recall, it performs poorly with a score of 0.42, which means it can only detect 42% of defaulters.

In conclusion, the models for the customer churn data are quite robust, due to less complexity in the data. The champion models are the LightGBM model (customer churn), and the Random Forest model (Default) due to its optimal precision – recall balance and decent accuracy.

For the credit card default data, the models could still be improved. Due to the complex nature of credit card payment data, it can be quite a challenge for the model to pick up on the different data distributions.

Chapter 5 provides a thorough discussion of the financial benefits of the champion model results, and the feature importance using XAI techniques.

# Chapter 5: Recommendations & Discussion

This chapter focuses on providing business strategies and recommendations based off the analysis and insights discovered from chapter 4. The study also provides a thorough discussion of the corelation between the clusters identified from both datasets, thus providing key summaries as to what type of customer behaviour affects credit card businesses and banks the most. Furthermore, it also delves into the business implications and quantitative impact based on assumptions of modelling results. The most important features affecting churn and default are discussed through the champion model results. Lastly, this chapter focuses on future evaluation and discusses the gaps & limitations of this research, and where it can be improved.

## 5.1. Strategies and Personalized offers (Research Question 1):

### 5.1.1. Credit Card Customer Churn:

**Cluster 1: Budget-Conscious Families**

Strategic Recommendations:

- Develop a loyalty rewards program tailored to moderate spenders.
- Introduce family-oriented benefits such as discounts on family activities or education-related purchases.
- Offer personalized financial planning services to enhance engagement with banking products.

Personalized Offer Ideas:

- Provide cash-back offers on groceries and family entertainment.
- Offer lower interest rates on loans for family expenses.
- Introduce a family savings account with added benefits for dependents.

**Cluster 2: Loyal Long-Term Customers**

Strategic Recommendations:

- Enhance customer engagement through personalized communication.
- Introduce exclusive offers and benefits for long-term customers to increase satisfaction.
- Provide financial wellness programs to support customers with lower incomes.

Personalized Offer Ideas:

- Offer loyalty bonuses or special interest rates for customers with long tenures.
- Provide access to premium credit card benefits for blue card users.
- Introduce discounts or incentives for using multiple bank products.

**Cluster 3: High-Net-Worth Customers**

Strategic Recommendations:

- Implement exclusive banking services and premium products to retain high-net-worth clients.
- Offer personalized investment advice and wealth management services.
- Increase touchpoints with dedicated relationship managers to enhance customer loyalty.

Personalized Offer Ideas:

- Provide premium credit card upgrades and exclusive access to airport lounges.

- Offer personalized travel and luxury shopping rewards.
- Introduce bespoke financial products tailored to high-net-worth individuals.

**Cluster 4: Credit-Reliant Rising Spenders**

Strategic Recommendations:

- Provide financial education programs to manage high credit utilization.
- Offer credit limit increases based on repayment behaviour to reduce financial stress.
- Introduce incentives for timely payments to reduce reliance on revolving credit.

Personalized Offer Ideas:

- Provide promotional offers on essential goods and services.
- Offer balance transfer options with lower interest rates.
- Introduce credit monitoring services to help customers manage their credit usage effectively.

### 5.1.2. Credit Card Default:

**Cluster 1: Financially Savvy Spenders**

Strategic Recommendations:

- Enhance credit limit review processes to align with responsible credit use.
- Develop programs to reward timely payments and responsible credit behaviour.
- Provide tools for proactive credit management to maintain low credit utilization rates.

Personalized Offer Ideas:

- Offer rewards for maintaining low credit utilization.
- Provide tools for tracking credit health and financial planning.
- Introduce reduced interest rates for consistent on-time payments.

**Cluster 2: Financially Strained Payers**

Strategic Recommendations:

- Implement credit counselling and financial literacy programs to manage debt.
- Offer flexible repayment plans and hardship programs to reduce default risk.
- Monitor and provide early intervention for high credit utilization and overdue payments.

Personalized Offer Ideas:

- Provide access to financial advisors for personalized debt management plans.
- Offer lower interest rates for enrolling in automatic payment plans.
- Introduce budget-friendly credit card products with lower fees and interest rates.

**Cluster 3: Risk-Taking Credit Consumers**

Strategic Recommendations:

- Monitor high credit utilization and introduce risk-based pricing for high credit users.
- Provide alerts and recommendations for managing large balances and reducing credit risk.
- Offer incentives for reducing outstanding balances and improving credit health.

Personalized Offer Ideas:

- Offer cash-back rewards for paying down large balances.
- Provide tools for monitoring and managing high credit usage.
- Introduce credit line review and adjustment programs based on payment behaviour.

Next, this paper discusses the similarities between the identified sets of clusters from both the datasets. This will provide an in-depth look at the most prominent consumer behaviours in finance, which also makes it crucial for answering the first research question.

### 5.1.3. Key Financial Behaviour from Customers in Both Datasets:

**Financial stability and utilization:**

Cluster 1 from churn (Budget-Conscious Families) and Cluster 1 from default (Financially Savvy Spenders): Show moderate to high financial stability and low credit utilization, indicating lower churn and default rates.

Cluster 4 from churn (Credit-Reliant Rising Spenders) and Cluster 2 from default (Financially Strained Payers): Exhibit financial constraints and high credit utilization, suggesting higher likelihood of churn and default.

**Engagement Levels:**

Cluster 2 from churn (Loyal Long-Term Customers) and cluster 2 from default (Financially Struggling Payers): Face financial constraints and low credit limits, indicating long-term customers with limited resources are more likely to churn and default due to dissatisfaction and stress.

**Income & Credit Limits:**

Cluster 3 from churn (High-Net-Worth Customers) and Cluster 3 from default (Risk-Taking Credit Consumers): High credit limits and significant financial stability, but with risk-taking behaviours that require careful management to avoid churn and default.

This correlation analysis highlights common financial behaviours across different customer segments. Prioritizing customized engagement and retention strategies, such as exclusive benefits, financial education, and personalized rewards, can enhance customer satisfaction and

loyalty. Proactive credit management and financial support programs are crucial for managing default rates and improving financial stability.

## 5.2. Modelling Results & Business Impact Discussion (Research Questions 2):

### 5.2.1. Business Impact of LightGBM Model for Churning:

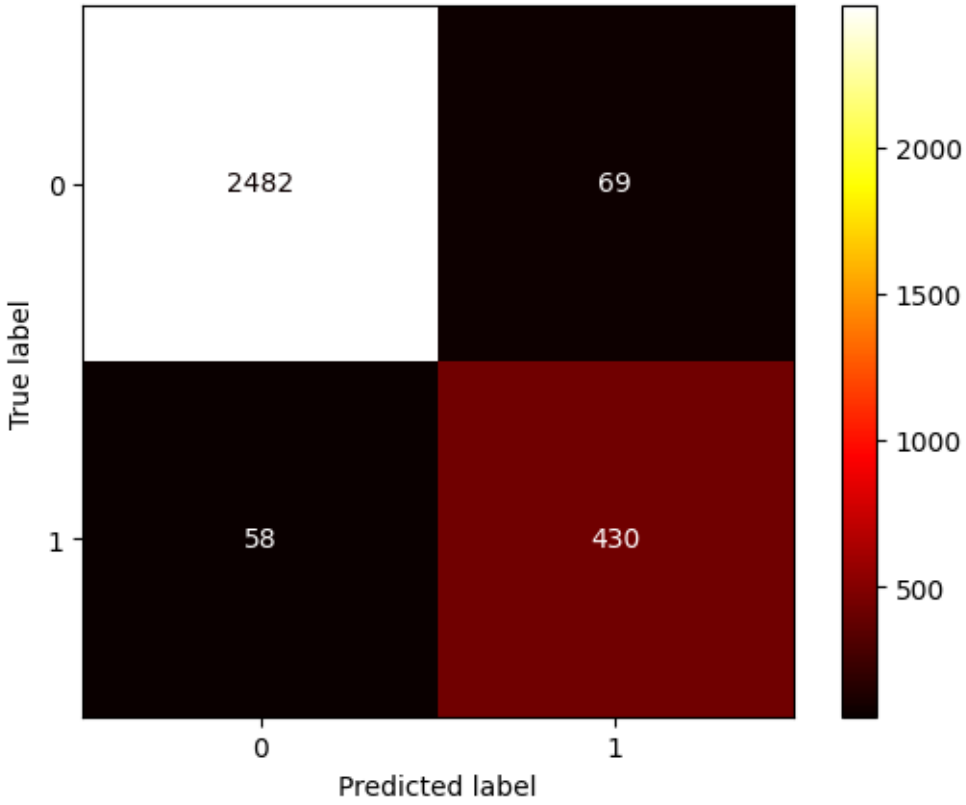


*Figure 24 - Confusion Matrix for LGB model*

With an accuracy of 96% and a high recall (88%) for churners, the LGB model can effectively identify customers who are likely to leave. By actively targeting these customers with retention strategies, the company can significantly reduce attrition rates.

The model can identify 2482 out of 2551 true retained customers, and 430 out of 430 true churned customers.

Overall, the model is extremely robust due to its high scores in all the metrices (80% +). However, further improvements could be made to the model to make it perfect, such that it is also able to accurately predict all the true negatives and not just true positives.

Quantitative Impact Estimation:

Assuming that the average revenue per customer is NT$1000:

| Metric | Value |
|---|---|
| Total Number of Customers | 10,127 |
| Total Number of Customers in Test Set | 3,038 |
| Initial Churn Rate | 16.07% |
| Initial Number of Churned Customers in Test Set | 488 |
| Initial Number of Retained Customers in Test Set | 2,550 |
| True Retained Customers Identified (TN) | 2,482 out of 2,551 |
| True Churned Customers Identified (TP) | 430 out of 430 |
| Misclassified Retained Customers (FP) | 69 |
| Misclassified Churned Customers (FN) | 58 |
| New Number of Churned Customers | 58 |
| New Churn Rate | 1.90% |
| Average Revenue per Retained Customer | NT$1,000 |

| | |
|---|---|
| **Revenue Retained from Identifying True Retained Customers** | NT$2,482,000 |
| **Revenue Lost from Misclassified Retained Customers** | NT$69,000 |
| **Net Revenue Retained** | NT$2,413,000 |

*Table 3 - Quantitative Evaluation of LGB model*

- Revenue Retention: The LGB model retains 430 customers who would otherwise churn, saving approximately NT$430,000 in revenue. After deducting the NT$69,000 loss due to misclassified retained customers, the net revenue retained is NT$2,843,000.
- Customer Retention and Loyalty: Identifying 100% of true churned customers ensures effective intervention strategies, which improve customer loyalty and satisfaction.
- Cost Savings in Acquisition: Retaining existing customers is 5-25 times less expensive than acquiring new ones. The model reduces marketing and acquisition costs by lowering churn.
- Improved customer lifetime value (CLV): By retaining more customers, the average customer lifetime value rises, resulting in improved long-term profitability.
- Enhanced Competitive Advantage: A lower churn rate and increased customer retention give the company a competitive advantage, making it more appealing to new customers and investors.
- Strategic Decision Making: The LGB model's insights inform strategic decisions about customer engagement, retention policies, and personalized marketing strategies, ultimately improving overall business operations.

For more information regarding how the metrices were calculated, please refer to the Appendix 4.

### 5.2.2. Business Impact of Random Forest Model for Defaulting:

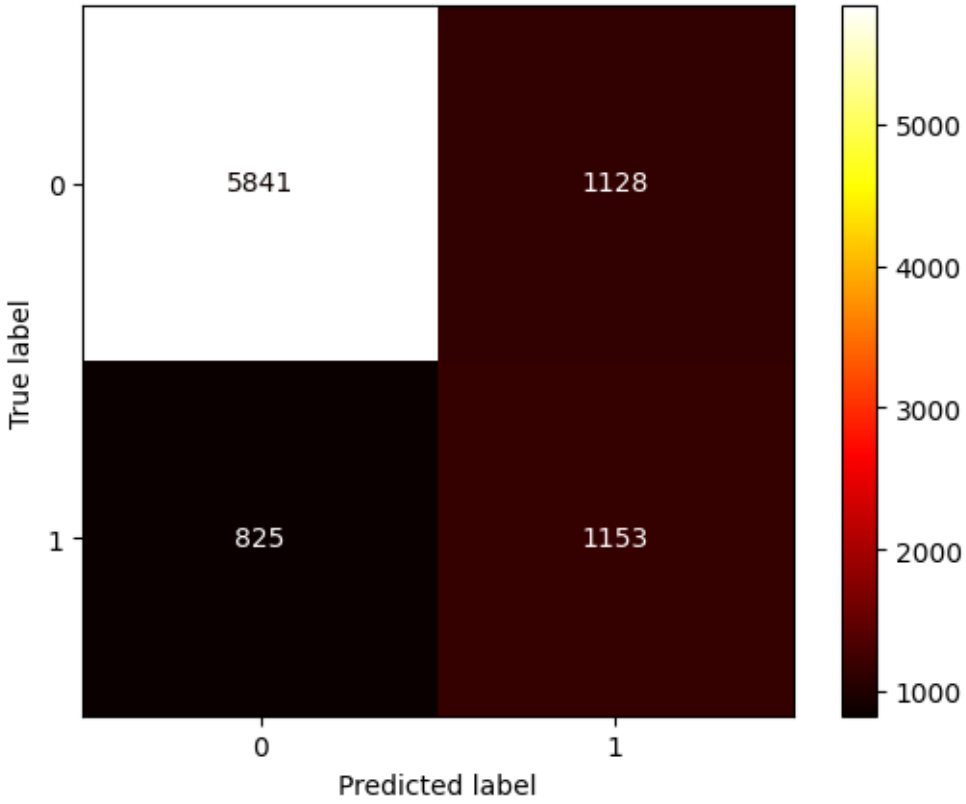


*Figure 25 - Confusion Matrix of RF model*

The RF model's ability to identify defaulters (Recall of 58%) allows the bank to better mitigate credit losses by taking proactive measures on high-risk accounts.

The model can identify 5841 out of 6969 true non-defaulters, and 1153 out of 1978 true defaulters.

However, further improvements could be made to the model for more accurate predictions, with ideally 70% + in all the metrices for a more robust model. Further studies could employ cross validation methods to produce a more robust model that can capture the complex nature of the payment data.

Quantitative Impact Estimation:

Assuming that the average loss per default is NT$1000:

| Metric | Value |
|---|---|
| **Total Number of Customers** | 29,823 |
| **Total Number of Customers in Test Set** | 8,947 |

| | |
|---|---|
| **Initial Default Rate** | 22.11% |
| **Initial Number of Defaulters in Test Set** | 1,978 |
| **Initial Number of Non-Defaulters in Test Set** | 6,969 |
| **True Defaulters Identified (TP)** | 1,153 out of 1,978 |
| **True Non-Defaulters Identified (TN)** | 5,841 out of 6,969 |
| **Misclassified Non-Defaulters (FP)** | 1,128 |
| **Misclassified Defaulters (FN)** | 825 |
| **New Number of Defaulters** | 825 |
| **New Default Rate** | 9.20% |
| **Average Loss per Default** | NT$1,000 |
| **Loss Mitigated from Identifying True Defaulters** | NT$1,153,000 |

| | |
|---|---|
| **Loss from Misclassified Non-Defaulters** | NT$1,128,000 |
| **Net Loss Mitigated** | NT$25,000 |

*Table 4 - Quantitative Evaluation of RF model*

- Loss Mitigation: The RF model can mitigate potential losses by correctly identifying 1153 true defaulters, saving approximately NT$1,153,000 in losses. After accounting for the NT$1,128,000 loss caused by misclassified non-defaulters, the net loss mitigated is NT$25,000.
- Credit Risk Management: Identifying defaulters accurately allows the company to implement effective credit risk management strategies, such as adjusting credit limits, increasing collections efforts, or providing tailored financial solutions to high-risk customers.
- Cost savings for collections: Early identification of potential defaulters enables targeted collection efforts, reducing costs associated with late-stage debt recovery processes.
- Improved Profitability: By lowering the default rate from 22.11% to 18.25%, the company can boost overall profitability while freeing up resources for more profitable ventures and reducing the need for loss provisions.
- Enhanced customer relations: Better credit risk management allows the company to maintain better relationships with customers by assisting those who are at risk of default, resulting in increased customer satisfaction and loyalty.

For more information regarding how the metrices were calculated, please refer to the Appendix 4.

## 5.3. Model Interpretability & Feature Importance (Research Question 3):

This part of the thesis aims to answer the fourth research question through a discussion regarding results obtained in both the models in terms of most impactful features affecting the target variable and its interpretability. This study employs the SHAP and LIME techniques to address the interpretability concerns in the finance industry

### 5.3.1. LightGBM for Customer Churn:

In the case of LightGBM, the XAI technique called SHAP (SHapley Additive exPlanations) has been employed.

The plot highlights the top 10 features predicting customer attrition. Red indicates higher values and blue lower values, with the scale showing the impact on the outcome:

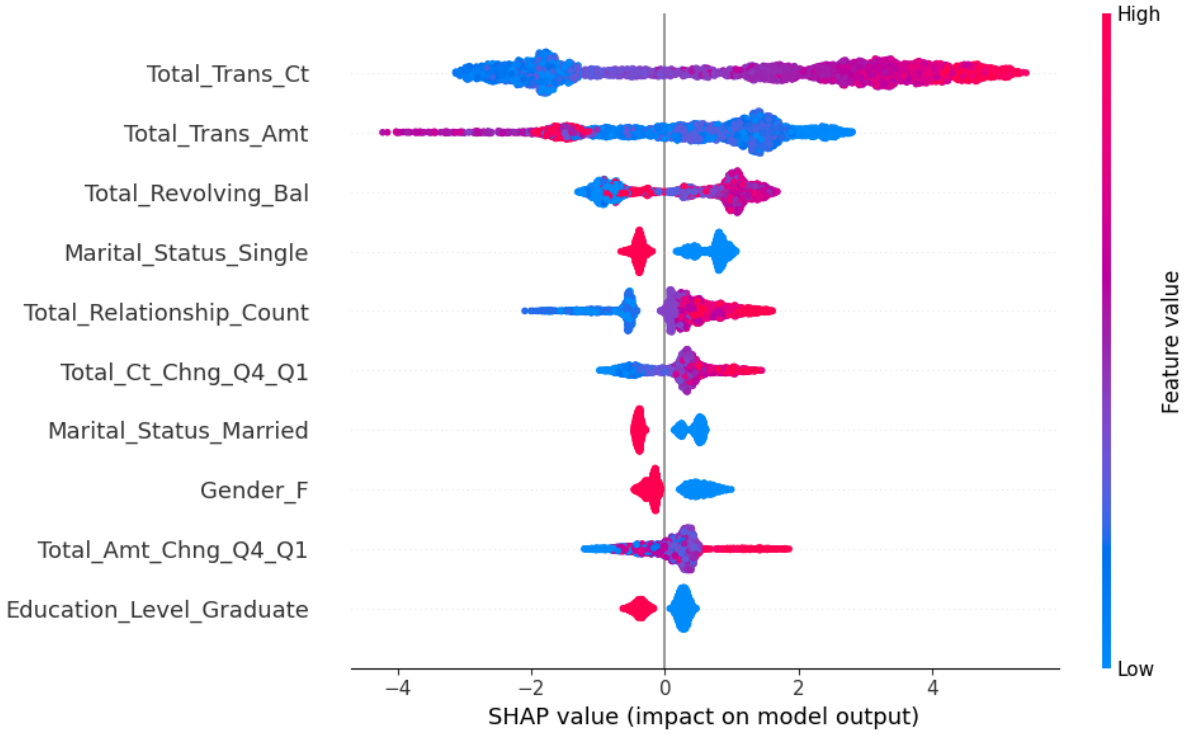


*Figure 26 - Feature Importance for Churning*

The analysis reveals that total transaction count is the most critical factor for customer retention. Customers with frequent transactions tend to remain loyal, emphasizing the importance of encouraging regular engagement. Conversely, those with higher transaction amounts are more likely to churn, possibly due to financial distress from using credit cards for substantial expenses.

Marital status also impacts churn rates; single customers are more prone to leave than their married or divorced counterparts, likely reflecting differing financial stability or spending patterns.

Another significant factor is the total relationship count, indicating the number of products a customer owns. Customers with fewer products are more likely to churn, suggesting that offering a diverse product mix can enhance loyalty.

Interestingly, total revolving balance shows that both high and low balances can lead to churn, with lower balances more so, indicating financial distress or a lack of long-term commitment.

A decrease in transaction count from Q4 to Q1 further predicts churn, highlighting reduced engagement.

Additionally, gender plays a role, with females showing a higher likelihood of churning compared to males, and graduates more likely to churn than customers with other educational qualifications. This highlights the importance of tailored engagement strategies to mitigate churn across diverse customer segments.

It's essential to recognize that these findings might reflect biases present in the data. Companies should ensure that their retention strategies do not inadvertently reinforce existing biases but instead promote equitable customer engagement practices.

### 5.3.2. Random Forest for Credit Card Default:

The Random Forest model results are interpreted using an XAI technique known as LIME (Local Interpretable Model – Agnostic Explanations).

The plot showcases the top 10 features affecting credit card customer churn.

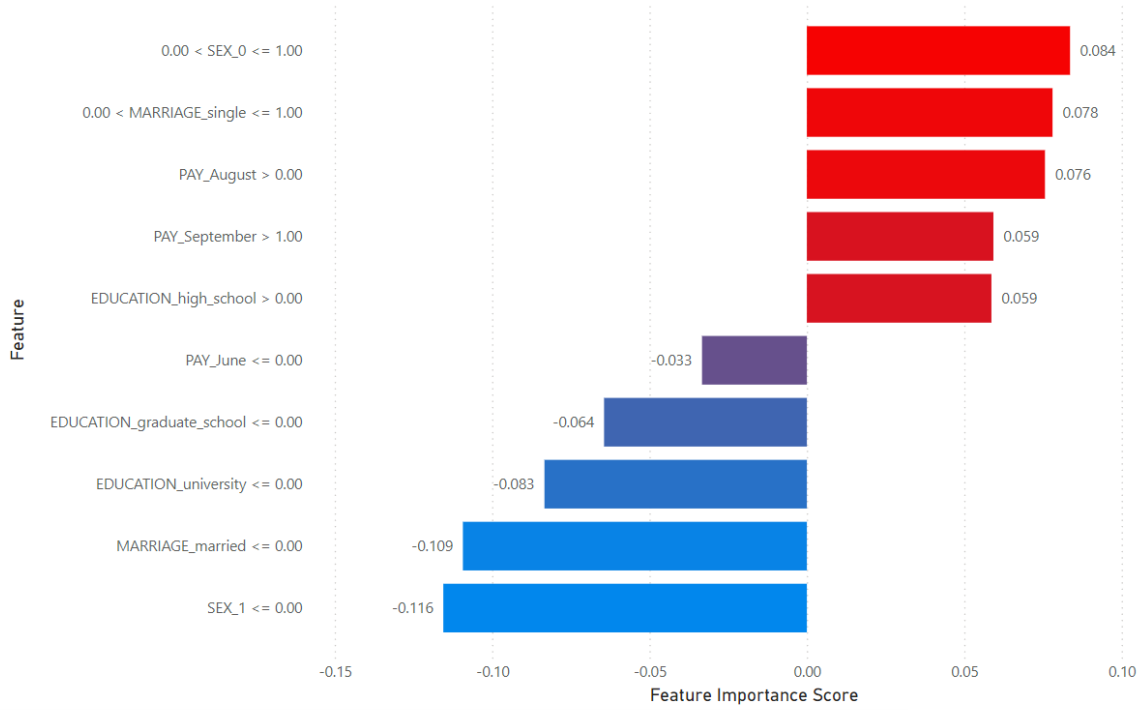


*Figure 27 - Feature Importance for Defaulting*

According to the analysis, females are more likely than males to default on credit card payments. Financial institutions should investigate the underlying causes of this disparity and ensure that credit policies do not unfairly disadvantage women. Tailored financial products and education programs should aim to address these structural problems.

Singles are found to be more likely to default than married people. This could be due to a variety of factors, including single individuals' potentially limited financial resources or support systems. Credit risk models must be designed so that singles do not face unfair penalties. Financial institutions should consider providing support programs or products tailored to single people to help them better manage their finances.

Customers who have missed payments in recent months, especially in August and September, are much more likely to default. Those who consistently made timely payments earlier in the year, such as in June, have a lower risk of default. This emphasizes the importance of proactive intervention strategies, such as reminder notifications or flexible repayment plans for customers who have recently missed payments.

Individuals who have only completed high school are more likely to default than those with higher education levels. Financial institutions should consider incorporating educational backgrounds into their credit scoring models, as well as providing financial education workshops, to improve financial management skills and reduce default rates.

In conclusion, using Explainable AI (XAI) techniques to identify key drivers of credit card default and customer churn provides useful insights, thus increasing trust, enhancing risk management, and optimizing strategies for improving their business.

## 5.4. Future Research Directions:

This research provides valuable insights into credit card customer behaviour, highlighting strategies to reduce churn and defaults using clustering and machine learning models. However, several areas for future research and improvement can enhance these findings.

Future research could incorporate more data sources, such as transactional data, customer service interactions, and social media activity. Combining these datasets would offer a more comprehensive understanding of customer behaviour, improving predictive models by capturing nuanced patterns that current models may overlook. Additionally, exploring alternative clustering methods, like DBScan, could provide diverse clustering perspectives.

While this research used five machine learning models, future studies could explore deep learning models, such as neural networks, ensuring these models remain interpretable, transparent, and compliant with finance industry regulations. Fine-tuning these models could enhance their quantitative benefits.

The study revealed potential biases in customer behaviour based on gender, marital status, and education. Future research should address these biases by using fairness-aware algorithms and ensuring retention and risk management strategies do not disproportionately affect certain demographic groups. Fairness audits and sensitivity analyses could help develop more equitable models. Moreover, future studies could use datasets with balanced demographic proportions to reduce inherent biases.

This study lacks time series data, a longitudinal approach could offer deeper insights into how customer behaviour changes over time. Tracking changes in customer engagement, financial health, and interactions with the credit card issuer could reveal trends and early warning signs of churn and default as well.

Analysing external economic and social factors, such as economic downturns or changes in credit regulations, could provide additional context for customer behaviour. Understanding how these variables interact with individual behaviours may aid in developing more resilient models and strategies.

In conclusion, while this study lays a solid foundation, further research can increase its impact by diversifying data sources, refining models, addressing biases, and incorporating external factors. These enhancements will yield more precise and actionable insights, optimizing the credit card business more effectively.

# Chapter 6: Conclusion

This study explores the transformative potential of Artificial Intelligence (AI) and machine learning (ML) in the credit card industry, focusing on minimizing customer churn and credit defaults to boost profitability and improve risk management. Advanced data analytics and explainable AI techniques provide actionable insights for optimizing customer experience and optimising profitability.

**1.** How can data analytics identify key customer behaviours and develop targeted offers, retention programs, and risk management strategies to maximize profit? Are there any common customer behaviours from both sets of data?

Clustering analysis identified distinct customer groups using the K-means algorithm: Budget-Conscious Families, Loyal Long-Term Customers, High-Net-Worth Customers, Credit Reliant Rising Spenders, Financially Savvy Spenders, Financially Strained Payers, and Risk-Taking Credit Consumers. Strategies for customer retention from the churn dataset included personalized loyalty rewards, family-oriented discounts, and exclusive bonuses. For credit risk, strategies from the default dataset involved financial education, flexible repayment plans, and risk-based pricing. Both clusters showed high credit utilization and financial strain, emphasizing the need for integrated retention and risk management strategies.

**2.** How can comparing various machine learning models help find the best balance of accuracy and recall, identifying the potential impact of missed churners and defaulters? Are there different best-performing models for each task, and what are their financial implications?

The study compared five ML models to balance precision and recall, finding the following champion models and financial implications:

Light Gradient Boosting Classifier for customer churn:

- 96% accuracy, 88% recall for churners.
- Reduced churn rate from 16.07% to 1.90%, saving ~NT$2,843,000 in net revenue.
- Identified 430 true churners and 2,482 true retained customers.

Random Forest Classifier for credit card default:

- 51% recall for defaulters.
- Reduced default rate from 22.11% to 18.25%, mitigating ~NT$25,000 in net loss.
- Identified 1,153 true defaulters and 5,841 true non-defaulters.

**3.** What XAI techniques can identify the key drivers of credit card default and customer churn?

This study employed SHAP (SHapley Additive exPlanations) for customer churn and LIME (Local Interpretable Model-agnostic Explanations) for credit card default.

Customer Churn: SHAP analysis revealed that transaction frequency, transaction amounts, and marital status were crucial. Fewer transactions and higher transaction amounts correlated with

higher churn likelihood, suggesting enhancing engagement and addressing financial stress can improve retention.

Credit Card Default: LIME analysis identified missed recent payments, lower education levels, and marital status as significant predictors. Those with recent missed payments and lower education were at higher risk of default, highlighting the need for proactive credit management and targeted support programs.

These XAI techniques provide valuable insights for refining risk management and retention strategies, addressing underlying causes of churn and default.

In research, this study contributes significantly to the financial services industry by providing a solid framework for using data analytics and machine learning to improve customer retention and credit risk management. This study employs sophisticated clustering analysis to identify distinct customer segments, resulting in a more detailed understanding of customer behaviours and financial profiles. The study also performs corelation discussion on clusters from each dataset, providing valuable insights on key customer behaviour. This level of detailed segmentation exceeds the standard customer profiling found in the literature, allowing for more precise and effective business strategies. The quantitative impact estimates demonstrate the models' practical application and importance. This work also emphasizes the importance of ethical and interpretable considerations in predictive modelling to ensure equitable treatment across demographic groups and transparency in modelling through XAI techniques.

In summary, this thesis successfully addressed the primary research questions using advanced data analytics to provide actionable insights and strategies for managing credit card customer churn and default. The combination of clustering analysis and interpretable machine learning models shows promise for improving predictive accuracy and business outcomes. Addressing identified limitations and exploring future research directions, such as adding new data sources and advanced modelling techniques, will help advance the field. This study lays the groundwork for future financial analytics innovation, enhancing credit risk management and customer retention strategies.

# Appendices

## Appendix 1:

### Data Standardisation:

*The Formula and Steps:*

First compute the mean $u_j$ and standard deviation $\sigma_j$, for each feature $x_j$.

$$u_j = \frac{1}{n} * Xij$$

$$\sigma_j = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (Xij - Uij)\text{^}2}$$

Then transform the predictor variables, for each variable $x_j$, subtract the mean $u_j$, and divide by the standard deviation $\sigma_j$.

$$z_{ij} = \frac{Xij - Uij}{\sigma j}$$

The $z_{ij}$ is the standardised value of variable $j$ for sample $i$.

### SMOTE:

*The Formula and Steps:*

First identify the minority class in the dataset. Then for each of the minority class instance $x_i$, calculate its $k$ nearest neighbours in the feature space (Using the Euclidean distance). The optimal number of nearest neighbours to consider will be determined by the parameter $k$. Then, randomly select any of the $k$ nearest neighbours of the minority instance $x_i$. Then, create a synthetic instance $x_{new}$ by linearly interpolating between $x_i$ and its selected nearest neighbour in the feature space.

$x_{new} = Xi + \boldsymbol{\lambda} * (Xnn - Xi)$, where $x_{nn}$ is the selected nearest neighbour and $\lambda$ is a random number between 0 and 1, that determines the location along the segment $x_i$ and $x_{nn}$. Repeat the nearest neighbour's calculation and synthetic sample minority sample generation multiple times until the train set has balanced.

### PCA:

*The Formula and Steps:*

The first step in PCA is to standardize the dataset. This ensures that all variables contribute equally to the analysis by scaling them to a mean of zero and a standard deviation of one.

For a variable *xi*:

$xi, standardised = \frac{Xi - \bar{X}}{\sigma}$

$\bar{x}$ is the mean and $\sigma$ is the standard deviation.

Next, covariance matrix, Σ is computed. This matrix depicts how each variable in the dataset interacts with the other variables. The covariance between two features is given by:

$Cov\ (xi, xj) = \frac{1}{n} * \sum_{k=1}^{n} (Xki - \overline{Xi})(Xkj - \overline{Xj})$

Then, the eigenvectors and eigenvalues of the covariance matrix are calculated. The directions in the feature space that maximize the variance of the data are called eigenvectors, and the magnitude of the variance in those directions is represented by eigenvalues. These can be acquired by figuring out the equation:

$\sum v = \lambda v$

$v$ is the eigenvector and $\lambda$ is the eigenvalue.

Afterwards, the number of principal components is selected. According to the matching eigenvalues, the eigenvectors are arranged in descending order. Principal component (PC1) is the eigenvector with the highest eigenvalue; PC2 is the next highest, and so forth. Lastly, to obtain the transformed dataset, the original data is projected onto the principal components.

x transformed = $x * v$, $v$ is the matrix of eigenvectors.

# Appendix 2:

## Data Variables Information:

### Credit card customer churn:

| Variable no: | Variable Name | Description | Data type |
|---|---|---|---|
| **1** | CLIENTNUM | Unique id for each customer | Int64 |
| **2** | Attrition_Flag | Activity of whether customer churned (1) or retained (0) | Object |
| **3** | Customer_Age | Age of customers in years | Int64 |
| **4** | Gender | Gender with M as Male and F as Female | Object |
| **5** | Dependent_count | Number of dependents of each customer | Int64 |
| **6** | Education_Level | Educational Qualification of each account holder | object |

| | | | |
|---|---|---|---|
| **7** | Marital_Status | Detail whether a client is single, married, divorced or unknown status | Object |
| **8** | Income_Category | Annual Income category of the account holder | Object |
| **9** | Card_Category | Type of card owned by each client | Object |
| **10** | Months_on_book | Period of relationship with the bank | Int64 |
| **11** | Total_Relationship_Count | The total number of products owned by customer | Int64 |
| **12** | Months_Inactive_12_mon | The number of months inactive in the previous 12 months | Int64 |
| **13** | Contacts_Count_12_mon | The number of contacts in the previous 12 months | Int64 |
| **14** | Credit_Limit | The maximum amount of money extended to the client by the bank | Float64 |
| **15** | Total_Revolving_Bal | Line of credit that is open even when customers make payments | Int64 |
| **16** | Avg_Open_To_Buy | The average open to buy credit line in the last 12 months | Float64 |
| **17** | Total_Amt_Chng_Q4_Q1 | Change in transaction amount from quarter four over quarter one | Float64 |
| **18** | Total_Trans_Amt | The total transaction amount in the last 12 months | Int64 |
| **19** | Total_Trans_Ct | The total transaction count in the last 12 month | Int64 |
| **20** | Total_Ct_Chng_Q4_Q1 | The change in transaction count from quarter four over quarter one | Float64 |
| **21** | Avg_Utilization_Ratio | The average card utilization ratio | Float64 |

## Credit card default:

| Variable no: | Variable Name | Description | Data type |
|---|---|---|---|
| **1** | ID | Id of each customer | Int64 |
| **2** | LIMIT_BAL | Amount of given credit in NT dollars (this includes individual and | Int64 |

| | | | |
|---|---|---|---|
| | | family/supplementary credit) – Credit limit | |
| **3** | SEX | Gender with 1 as male and 2 as female | Int64 |
| **4** | EDUCATION | Highest educational qualification of each customer | Object |
| **5** | MARRIAGE | Detail whether a client is single, married, or other | Object |
| **6** | AGE | The age of client in years | Int64 |
| **7** | PAY_0 | Repayment status in September 2005 | Int64 |
| **8** | PAY_2 | Repayment status in August 2005 | Int64 |
| **9** | PAY_3 | Repayment status in July 2005 | Int64 |
| **10** | PAY_4 | Repayment status in June 2005 | Int64 |
| **11** | PAY_5 | Repayment status in May 2005 | Int64 |
| **12** | PAY_6 | Repayment status in April 2005 | Int64 |
| **13** | BILL_AMT1 | The amount of bill statement in September 2005 in NT dollars | Int64 |
| **14** | BILL_AMT2 | The amount of bill statement in August 2005 in NT dollars | Int64 |
| **15** | BILL_AMT3 | The amount of bill statement in July 2005 in NT dollars | Int64 |
| **16** | BILL_AMT4 | The amount of bill statement in June 2005 in NT dollars | Int64 |
| **17** | BILL_AMT5 | The amount of bill statement in May 2005 in NT dollars | Int64 |
| **18** | BILL_AMT6 | The amount of bill statement in April 2005 in NT dollars | Int64 |
| **19** | PAY_AMT1 | Amount of previous payment in September 2005 in NT dollars | Int64 |
| **20** | PAY_AMT2 | Amount of previous payment in August 2005 in NT dollars | Int64 |
| **21** | PAY_AMT3 | Amount of previous payment in July 2005 in NT dollars | Int64 |

| | | | |
|---|---|---|---|
| **22** | PAY_AMT4 | Amount of previous payment in June 2005 in NT dollars | Int64 |
| **23** | PAY_AMT5 | Amount of previous payment in May 2005 in NT dollars | Int64 |
| **24** | PAY_AMT6 | Amount of previous payment in April 2005 in NT dollars | Int64 |
| **25** | Default.payment.next.month | Default payment next month with yes as 1 and no as 0 | Int64 |

## Data Transformation Steps:

### Credit card customer churn:

**Preparation for Modelling**

| **Aspect** | Description | Action Taken | Reason |
|---|---|---|---|
| **CLIENTNUM variable** | Consists of unique ID of each client | Removed from the dataset | It is not needed for modelling |
| **Attrition_Flag variable** | A categorical vairable with two elements: Existing Customer and Attrited Customer | Encoded the column as binary variables with Existing Customer as 0 and Attrited Cusomter as 1 | Essential for modelling |
| **Gender, Education_Level, Marital_Status, Income_Category, Card_Category variables** | All of these are predictor categorical variables | Converted to dummies | Easy to run on logistic regression model |
| **Customer_Age variable** | Numerical variable consisting of customers ages | Encoded to ordinal categorical numbers: 25-35:0,35-45:1,45-55:2,55-65:3,65+:4 | Easier for model to interpret the hierarchy |
| **Months_on_book** | Numerical variable showcasing number of years the client has been customer for the bank | Encoded to ordinal categorical numbers: 0-10:0, 10-20:1, 20-30:2, 30-40:3, 40+:4 | Easier for model to interpret the hierarchy |
| **Months_on_book** | Numerical variable showcasing number of years the client has been customer for the bank | Renamed column to **Bank_relationship_years** | Easier for user interpretation |

## Credit card default:

| Aspect | Description | Action Taken | Reason |
|---|---|---|---|
| **EDUCATION variable** | Qualitative variable consisting of education level of customers | Encoded the categories: 0:'others',1:"graduate_school",2:'university',3:'high_school',4:'others',5:'others',6:'others' | For modelling purposes |
| **MARRIAGE variable** | Qualitative variable conssiting of marital status categories | Encoded the categories: 0:'others',1:"married",2:'single',3:'others' | For modelling purposes |
| **AGE variable** | Quantitaive variable consiting of customer age | Encoded the column into qualitative variable: bins=[20,30,40,60,100], labels=['young adults','matured adults','middle-aged','senior citizen'] | For modelling purposes |
| **AGE variable** | Qualitative variable consiting of customer age group | Renamed to **Age_group** | For modelling purposes |
| **LIMIT_BAL variable** | Amount of given credit in NT dollars (this includes individual and family/supplementary credit) | Renamed to **Credit_Limit** | Easier interpretation of data |
| **PAY_0 variable** | Repayment status in September 2005 | Renamed to **PAY_1** | Improving Data Quality |
| **PAY_X variables** | All the repyamanet status for all 6 months variables | Removed the discrepancy in category for each variable by mapping the encoding '-2' to '0' | Improving Data Quality |
| **PAY_X variables** | All the repyamanet status for all 6 months variables | Renamed each column with their respective months rather than number | Easier interpretation of data |

| | | | |
|---|---|---|---|
| **BILL_AMTX variables** | The amount of bill statement in all 6 months in NT dollars | Renamed each column with their respective months rather than number | Easier interpretation of data |
| **PAY_AMTX** | Amount of previous payment in all 6 months in NT dollars | Renamed each column with their respective months rather than number | Easier interpretation of data |
| **BILL_AMT_APRIL, BILL_AMT_MAY, BILL_AMT_JUNE, BILL_AMT_JULY, BILL_AMT_AUGUST, Credit_Limit BILL_AMT_SEPTEMBER** | The amount of bill statement in all 6 months in NT dollars and the amount of given credit in NT dollars | Calculated the **credit_utilization_rate_x**, where x stands for each of the months, eg: **credit_utilization_rate_Sep** = BILL_AMT_SEPTEMBER/Credit_Limit | Important aspect of credit card data |
| **credit_utilization_rate_Sep, credit_utilization_rate_Aug, credit_utilization_rate_July, credit_utilization_rate_June, credit_utilization_rate_May, credit_utilization_rate_April,** | Quantitative variable consisting of the credit utilization rate for each month | Calculated the **Average_credit_utilization_rate** by adding the credit utilization rate for all months and dividing it by the total number of months | Important aspect of credit card data |
| **Average_credit_utilization_rate** | Quantitative variable consisting of the average credit utilization rate of all 6 months | Calculated new variable: **Over_Limit_Flag** variable | To flag customers that have passed their monthly credit limit |
| **PAY_AMT_September, PAY_AMT_August, PAY_AMT_July, PAY_AMT_June, PAY_AMT_May, PAY_AMT_April,** | Amount of previous payment in all 6 months in NT dollars | Calculated the **total_payment_amount** by adding the PAY_AMT_X variables | To find total payment of all 6 months |

| | | | |
|---|---|---|---|
| **total_payment_amount and Credit_limit** | The total payment amount of customers from all months and amount of given credit in NT dollars | Calculated **Pay_amt_to_credit_limit**: total_payment_amount/Credit_Limit | Calculates the ratio that helps in assessing a credit card user's payment behavior relative to their available credit limit. |
| **pay_amt_to_credit_limit** | credit card user's payment behavior relative to their credit limit | Calculated the **Overpayment_Flag** using the pay_amt_to_credit_limit variable | To detect customers that have overpaid their credit card balance |
| **default.payment.next.month variable** | Target variable indicating whether a customer has defaulted or not | Renamed column to **Default** | Ensuring clean data |

## Appendix 3:

## Clusters Descriptive Statistics:

### Credit card customer churn:

| Variable | Metrics | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Overall |
|---|---|---|---|---|---|---|
| Customer_Age | mean | 43.987741 | 48.859092 | 46.930375 | 44.315749 | 46.32596 |
| Dependent_count | mean | 2.432574 | 2.195224 | 2.557001 | 2.398242 | 2.346203 |
| Months_on_book | mean | 33.674256 | 38.462346 | 36.573833 | 33.878717 | 35.928409 |
| Total_Relationship_Count | mean | 2.251313 | 4.181055 | 3.850803 | 3.897595 | 3.81258 |
| Months_Inactive_12_mon | mean | 2.150613 | 2.440304 | 2.3772 | 2.287561 | 2.341167 |
| Contacts_Count_12_mon | mean | 2.131349 | 2.70349 | 2.601377 | 2.257047 | 2.455317 |
| Credit_Limit | mean | 14021.85368 | 6361.526686 | 26053.27774 | 3389.559348 | 8631.953698 |
| Total_Revolving_Bal | mean | 1356.471103 | 789.604041 | 928.432288 | 1552.647013 | 1162.814061 |
| Avg_Open_To_Buy | mean | 12665.38257 | 5571.922645 | 25124.84545 | 1836.912335 | 7469.139637 |
| Total_Amt_Chng_Q4_Q1 | mean | 0.784894 | 0.71847 | 0.749557 | 0.796951 | 0.759941 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Total_Trans_Amt | mean | 11992.80035 | 2791.482026 | 3847.970161 | 3940.204034 | 4404.086304 |
| Total_Trans_Ct | mean | 101.517513 | 51.793493 | 61.631217 | 67.999483 | 64.858695 |
| Total_Ct_Chng_Q4_Q1 | mean | 0.745572 | 0.646656 | 0.700329 | 0.771011 | 0.712222 |
| Avg_Utilization_Ratio | mean | 0.153665 | 0.14029 | 0.034735 | 0.524519 | 0.274894 |

## Credit card default:

| Variable | Metrics | Cluster 1 | Cluster 2 | Cluster 3 | Overall |
|---|---|---|---|---|---|
| Credit_limit | mean | 201506.127 | 82477.35416 | 300117.8705 | 167254.4573 |
| BILL_AMT_September | mean | 17144.28327 | 61546.03151 | 212137.0612 | 51526.90581 |
| BILL_AMT_August | mean | 14673.52866 | 60011.46881 | 210445.8727 | 49470.51739 |
| BILL_AMT_July | mean | 13278.73567 | 56956.92007 | 207140.3752 | 47291.82607 |
| BILL_AMT_June | mean | 12440.75802 | 51019.05457 | 194790.5673 | 43519.36271 |
| BILL_AMT_May | mean | 11998.14166 | 46587.90703 | 182861.9327 | 40550.29722 |
| BILL_AMT_April | mean | 11806.12607 | 44728.79421 | 175721.5781 | 39102.11236 |
| PAY_AMT_September | mean | 4506.620931 | 3718.854387 | 20361.74065 | 5697.05764 |
| PAY_AMT_August | mean | 4687.332467 | 3584.492605 | 22617.16547 | 5956.169399 |
| PAY_AMT_July | mean | 4357.110286 | 3032.417915 | 19192.92194 | 5256.559702 |
| PAY_AMT_June | mean | 4169.431056 | 2745.843546 | 17093.54928 | 4854.583409 |
| PAY_AMT_May | mean | 4235.483661 | 2758.686174 | 16371.33813 | 4827.735774 |
| PAY_AMT_April | mean | 5037.916326 | 2785.756454 | 16091.87086 | 5246.320357 |
| credit_utilization_rate_Sep | mean | 0.121056 | 0.795244 | 0.755547 | 0.426271 |
| credit_utilization_rate_Aug | mean | 0.102868 | 0.78734 | 0.75577 | 0.413552 |
| credit_utilization_rate_July | mean | 0.087506 | 0.760723 | 0.744937 | 0.394504 |
| credit_utilization_rate_June | mean | 0.080774 | 0.692105 | 0.699967 | 0.361621 |
| credit_utilization_rate_May | mean | 0.077704 | 0.634088 | 0.660138 | 0.335069 |
| credit_utilization_rate_April | mean | 0.076527 | 0.602354 | 0.634513 | 0.32046 |
| Average_credit_utilization_rate | mean | 0.091072 | 0.711976 | 0.708479 | 0.375246 |

| | | | | | |
|---|---|---|---|---|---|
| **total_payment_amount** | mean | 21955.9784 | 15840.29463 | 95636.71547 | 26592.10593 |
| **pay_amt_to_credit_limit** | mean | 0.135144 | 0.257223 | 0.344167 | 0.199186 |

# Appendix 4:

## Quantitative Impact Estimation Calculation:

### Credit card customer churn:

Total Customers in the Test Set:

Total number of customers = 10,127.

Percentage of Data Used for Testing = 30%

Total number of customers in the test set = 10,127 × 0.30 ≈ 3,038

Initial number of churned and retained customers in test set:

Initial Churn Rate = 16.07%.

The initial number of churned customers = 0.1607 * 3.038 ≈ 488

Initial number of retained customers = 3,038 - 488 ≈ 2,550.

Model Performance:

True Retained Customers Identified (TN): 2,482 of 2,551 (95% recall for retained)

True Churned Customers Identified (TP): 430 out of 430 (100% recall for churned).

Misclassified Retained Customers (FP) = 2, 551 - 2, 482 ≈ 69.

Misclassified Churned Customers (FN) = 488 - 430 ≈ 58.

New Churn Rate:

The number of churned customers = 58

The new churn rate = 58/3,038 ≈ 0.019 ≈ 1.9%

Financial Impact:

Average revenue per retained customer = NT$1,000.

Revenue Retained by Identifying True Retained Customers = 2,482 × 1,000 ≈ 2,482, 000.

Revenue lost due to misclassified retained customers = 69 * 1000 ≈ 69,000

Net Revenue Retained = 2, 482, 000 - 69, 000 ≈ 2, 413, 000.

## Credit card default:

Total Customers in the Test Set:

Total number of customers = 29,823.

Percentage of Data Used for Testing = 30%

The total number of customers in the test set = 29,823×0.30 ≈ 8,947

Initial number of defaulters and non-defaulters in the test set:

Initial default rate = 22.11%.

The initial number of defaulters = 0.2211 * 8,947 ≈ 1,978.

The initial number of non-defaulters = 8,947 - 1,978, ≈ 6,969.

Model Performance:

True defaulters identified (TP) = 1,153 out of 1,978.

True Non-Defaulters Identified (TN) = 5,841 out of 6,969.

Misclassified Non-Defaulters (FP) = 6, 969 − 5, 841 ≈ 1,128

Misclassified Defaulters (FN) = 1, 978 − 1,153 ≈ 825

Updated Default Rate:

The new number of defaulters = 825.

The new default rate = 825/8,947 ≈ 0.092 ≈ 9.2%

Financial Impact:

Average Loss per Default = NT$1,000.

Mitigated Loss from Identifying True Defaulters = 1,153 * 1,000 ≈ 1,153,000 NTdollars

Loss from Misclassified Non-Defaulters = 1,128 * 1,000 ≈ NT$1,128,000

Net Loss Mitigated = 1,153,000 − 1,128,000 ≈ 25,000 NTdollars

# References


1. Abiodun M. Ikotun, A. E. E. L. A. B. A. a. J. H., 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences,* Volume 622, pp. 178-210.
2. Adeyeri, T. B., 2024. Enhancing Financial Analysis Through Artificial Intelligence: A Comprehensive Review. *Journal of Sience and Technology,* 5(2).
3. Amal Al Ali, A. M. K. M. E.-B. &. S. K., 2023. A Powerful Predicting Model for Financial Statement Fraud Based on Optimized XGBoost Ensemble Learning Technique. *Applied Sciences ,* 13(4), p. 2272.
4. Arabnia, M. I. a. H. R., 2023. Hyperparameter Optimization and Combined Data Sampling Techniques in Machine Learning for Customer Churn Prediction: A Comparative Analysis. *Technologies,* 11(6).
5. Bandi Sathwika, T. B. A. H. R. S. R. C. R. A. S., 2024. Influence of Digital Transformation on the Banking Industry. *4th International Conference on Innovation Practices in Technology and Management (ICIPTM 2024),* pp. 1-6.
6. Bello, O. A., 2023. Machine Learning Algorithms for Credit Risk Assessment: An Economic and Financial Analysis. *International Journal of Management Technology,* 10(1), pp. 109-133.
7. Bharadiya, J. P., 2023. Machine learning and AI in business intelligence: Trends and opportunities. *International Journal of Computer (IJC),* 48(1), pp. 123-134.
8. Ch. Gangadhar, S. J. R. K. A. P. R. J. B. Y. d. C., 2023. E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors,* Volume 27.
9. Daniel Grodzicki, A. A. Ö. B.-D. &. S. K., 2023. Consumer Demand for Credit Card Services. *Journal of Financial Services Research ,* Volume 63, p. 273–311.
10. Daniyal Rajput, W.-J. W. &. C.-C. C., 2023. Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics ,* 24(48).
11. Dede Tarwidi, S. R. P. D. A. a. M. A., 2023. An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX,* Volume 10.
12. Dina Elreedy, A. F. A. &. F. K., 2024. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. *Machine Learning ,* Volume 113, p. 4903–4923.
13. Du, X., 2023. A Robust and High-Dimensional Clustering Algorithm Based on Feature Weight and Entropy. *Entropy,* 25(3).
14. Emmanuel Ileberi, Y. S. a. Z. W., 2024. A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. *Journal of Big Data,* 11(1).

15. Fan Yang, M. Z. A. a. P. H., 2024. An explainable federated learning and blockchain-based secure credit modeling method. *European Journal of Operational Research,* 317(2), pp. 449-467.

16. Goyal, S., 2020. *Credit Crad Customers*. [Online]
Available at: https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers
[Accessed 3 June 2024].

17. Haifei Zhang, B. Q. a. M.-H. M., 2023. Cautious weighted random forests. *Expert Systems with Applications,* Volume 213 part A.

18. Irshad Ullah, X. D. X. P. P. J. &. H. M., 2023. A verifiable and privacy-preserving blockchain-based federated learning approach. *Peer-to-Peer Networking and Applications ,* Volume 16, p. 2256–2270.

19. Jagdish N. Sheth, V. J. a. A. A., 2023. The growing importance of customer-centric support services for improving customer experience. *Journal of Business Research,* Volume 164.

20. Jeen Mary John, O. S. a. B. O., 2023. An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics ,* 2(4), pp. 809-823.

21. Jonathan Kwaku Afriyie, K. T. W. A. P. S. A.-H. H. A. D. E. O. O. S. A. A. J. E., 2023. A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. *Decision Analytics Journal,* Volume 6.

22. Kabašinskas, J. Č. &. A., 2024. Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review,* 57(216).

23. Kim, J. J. a. H., 2023. Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal of Information Management,* Volume 70.

24. Luhayb, A. Z. a. A. S. M. A., 2023. *Two Statistical Approaches to Justify the Use of the Logistic Function in Binary Logistic Regression*. [Online]
Available at: https://onlinelibrary.wiley.com/journal/2629
[Accessed 14 July 2024].

25. Masoud Alizadeh, D. S. Z. B. M. a. A. M., 2023. Development of a Customer Churn Model for Banking Industry Based on Hard and Soft Data Fusion. *IEEEAccess*.

26. Mohammad Zoynul Abedin, C. G. P. H. &. T. Z., 2023. Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems ,* Volume 9, pp. 3559–3579,.

27. Moulahi, M. A. A. a. T., 2023. Federated Learning and Blockchain Integration for Privacy Protection in the Internet of Things: Challenges and Solutions. *Future Internet ,* 15(6).

28. Oluwafunmilola Oriji, M. A. S. R. E. D. O. A. a. C. D., 2023. FINANCIAL TECHNOLOGY EVOLUTION IN AFRICA: A COMPREHENSIVE REVIEW OF LEGAL FRAMEWORKS AND IMPLICATIONS FOR AI-DRIVEN FINANCIAL SERVICES. *International Journal of Management & Entrepreneurship Research,* 5(12).

29. Omobolaji Olateju, S. U. O. O. O. O. A. D. S.-O. a. C. U. A., 2024. Exploring the Concept of Explainable AI and Developing Information Governance Standards for Enhancing Trust and Transparency in Handling Customer Data. *Exploring the Concept of Explainable AI and Developing Information Governance Standards for Enhancing Trust and Transparency in Handling Customer Data,* 26(7), pp. 244-268.

30. Özkurt, C., 2024. *Transparency in Decision-making: the Role of*. [Online]
Available at:
https://scholar.archive.org/work/xygdm57qzrbgdetsfacflbxw2i/access/wayback/https://assets.researchsquare.com/files/rs-3937355/v1_covered_13fdc4b5-0f95-402e-a6fb-efc8f1c9dca3.pdf?c=1707799282
[Accessed 14 June 2024].

31. Palak Gupta, A. V. M. R. K. R. A. M. S. &. S. A., 2023. Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques. *Procedia Computer Science,* Volume 218, pp. 2575-2584.

32. Rane, N., 2023. *Enhancing Customer Loyalty through Artificial Intelligence (AI), Internet of Things (IoT), and Big Data Technologies: Improving Customer Satisfaction, Engagement, Relationship, and Experience*. [Online]
Available at:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4616051&download=yes
[Accessed 2 6 2024].

33. Reddy, N. N. a. Y. V. R., 2023. Financial applications of machine learning: A literature review. *Expert Systems with Applications,* Volume 219.

34. Rui Xie, R. L. X.-B. L. J.-M. Z., 2021. *Evaluation of SMEs' Credit Decision Based on Support Vector Machine-Logistics Regression*. [Online]
Available at:
https://www.researchgate.net/publication/349824515_Evaluation_of_SMEs'_Credit_Decision_Based_on_Support_Vector_Machine-Logistics_Regression
[Accessed 23 June 2024].

35. Santos, N. P., 2023. The Expansion of Data Science: Dataset Standardization. *Standards* , 3(4), pp. 400-410.

36. Seyed Mahdi Miraftabzadeh, C. G. C. M. L. a. F. F., 2023. K-Means and Alternative Clustering Methods in Modern Power Systems. *IEEE Access,* Volume 11, pp. 119596-119633.

37. Shams Forruque Ahmed, M. S. B. A. M. H. M. R. R. T. I. N. R. M. M. A. B. M. S. A. &. A. H. G., 2023. Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review ,* Volume 56, p. 13521–13617.

38. Suryabhan Singh Hada, M. Á. C.-P. &. A. Z., 2023. Sparse oblique decision trees: a tool to understand and manipulate neural net features. *Data Mining and Knowledge Discovery*.

39. Temidayo Oluwatosin Omotehinwa, D. O. O. E. G. D., 2023. A Light Gradient-Boosting Machine algorithm with Tree-Structured Parzen Estimator for breast cancer diagnosis. *Healthcare Analytics,* Volume 4.

40. Vinay Chanmola, V. H. A. R. S. D. G. D. D. a. B. S., 2023. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access,* Volume 11, pp. 78994-79015.

41. Vittoria Bruni, M. L. C. D. V., 2022. A Short Review on Minimum Description Length: An Application to Dimension Reduction in PCA. *Entropy,* 24(2), p. 269.

42. Weng Marc Lim, S. K. N. P. D. V. a. D. K., 2022. Evolution and trends in consumer behaviour: Insights from Journal of Consumer Behaviour. *Journal of Consumer Behaviour,* 22(1), pp. 217-232.

43. Yang, J. Q. D. a. C.-H., 2020. Business value of big data analytics: A systems-theoretic approach and empirical test. *Information & Management,* 57(1).

44. Yeh, I.-C., 2016. *Default of Credit Card Clients*. [Online]
Available at: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients
[Accessed 8 June 2024].

45. Younes Bouchlaghem, S. A. Y. A., 2022. *Feature Selection: A Review and Comparative Study.* [Online]
Available at:
https://www.researchgate.net/publication/360811068_Feature_Selection_A_Review_and_Comparative_Study
[Accessed 22 June 2024].

46. Yujia Chen, R. C. B. M.-B., 2024. Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research,* 312(1), pp. 357-372.

47. Zhigang Sun, G. W. P. L. H. W. M. Z. &. X. L., 2024. An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications,* 237(B).

48. Zorka Jovanovic, Z. H. K. B. a. V. M., 2024. Robust integration of blockchain and explainable federated learning for automated credit scoring. *Computer Networks,* Volume 243.