Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A machine learning framework for customer value identification based on transaction behaviour

Raihan Sharif

raihan.sharif@student.usm.my

Universiti Sains Malaysia

Salmi Mohd Isa

Universiti Sains Malaysia

Ashik Choudhury

Jahangirnagar University

Mohammad Nasir Hossian

Comilla University

Research Article

Keywords:

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11300280/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

This research focuses on the critical task of identifying high-value customers by analysing the spending behaviours of customers of **Sprocket Central Pty Ltd** through the application of advanced machine learning techniques on the dataset collected dataset obtained from the KPMG Data Analytics Consulting Virtual Internship, accessed via Kaggle. This comprehensive dataset comprises detailed consumer transactional records, including monetary expenses, purchase frequency, recency of transactions, and demographic variables and so on. So, this study implements two widely used classification algorithms here to find their effectiveness which are **Random Forest Classifier** and **Logistic Regression model**. These models were selected to represent contrasting approaches in predictive analytics of the collected data, with Random Forest exemplifying ensemble-based along with non-linear modelling and Logistic Regression representing a traditional as well as interpretable linear method used by most of the analysts.

The objective of this research was to predict which customers fall into the high spender category that is pivotal for maximizing profitability and optimizing marketing efforts by **Sprocket Central Pty Ltd**. Model performance was rigorously evaluated using a suite of standard classification metrics, including accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC AUC) to understand the efficiency of model performance as well as the expected output from the dataset analysis. The Random Forest model consistently outperformed Logistic Regression across all metrics, achieving an AUC score of 0.87, signifying excellent discriminatory power between high and low spenders from the dataset. In contrast, Logistic Regression yielded a respectable but comparatively lower AUC of 0.81 which is reflecting limitations in capturing the complex, non-linear relationships inherent in consumer spending behaviour in constantly changing business environment.

Here, the feature importance analysis within the Random Forest model revealed that monetary value of past transactions is the most influential predictor, followed by transaction frequency, customer age, and recency of purchase. This insight confirms that both behavioural and demographic factors play a crucial role in defining customer value for business to ensure their efforts to be put in the right place. Furthermore, the challenges that has been encountered throughout the research such as data imbalance, where high spenders constitute a minority of the population and addresses these through careful metric selection rather than oversampling techniques. This shows a future scope for further research aligning this topic.

Overall, the research underscores the substantial benefits of employing ensemble machine learning approaches in customer segmentation tasks through offering similar businesses an enhanced predictive accuracy and more granular customer insights. It also supports the integration of such models into business intelligence and customer relationship management systems to facilitate targeted marketing campaigns, efficient resource allocation, and improved customer retention strategies which can be effectively done through similar dataset analysis. The findings contribute to the growing body of knowledge on machine learning applications in marketing analytics and provide a foundation for future

research to refining model robustness, interpretability, and deployment in dynamic retail environments by businesses.

1. Introduction

1.1 Background of the Study

In this era of data-driven decision-making, identifying high-value customers has become very essential for effective marketing and customer relationship management for retail businesses. Throughout their operations, businesses often accumulate vast amounts of transactional and behavioural data about their customers which provides an opportunity to apply Machine Learning techniques to extract actionable insights from those datasets. Traditional statistical methods often fall short in capturing complex patterns within the dataset to capture the actual scenario in the purchase patterns. Thus, this study has used machine learning models, specifically Random Forest and Logistic Regression classifiers, to predict customer spending behaviour and identify high-value segments to justify the efficiency of advanced machine learning model between stated models.

1.2 Problem Statement

According to Breiman (2001), businesses often face challenges in distinguishing high-value customers early in their engagement which causes problems in ensuring their smooth growth in their business. Relying on intuition or basic historical patterns may result in misclassification which might lead to inefficient resource allocation by retail businesses in the competitive market. Traditional logistic models which have been used from an early period, although interpretable, may not perform optimally with non-linear and high-dimensional data that are now collected by businesses from dynamic environments. Therefore, there is a need to assess whether advanced ML models like Random Forest can offer improved prediction accuracy and deeper behavioural insights for high-value customer identification from transactional datasets.

1.3 Aim

The aim of this research is *“to evaluate the effectiveness of machine learning models in identifying high-value customers based on their spending behaviour and other supporting information, using Random Forest and Logistic Regression classifiers”.*

1.4 Research Objectives

The objective of this research can be broken down into the following milestones:

- To analyze customer transactional data to determine spending patterns from the collected dataset.
- To build and train classification models using Python to get insights.
- To compare model performances using precision, recall, F1-score, and ROC AUC.

- To assess which model provides better identification of high-value customers to see the comparison along with the results using the efficient model.

1.5 Research Questions

Based on the objectives, the following questions can help to find precise answers to reach the expected point of this research:

- How effectively can machine learning models predict high-value customer behaviour?
- Which model performs better in identifying high spenders in this case?
- What performance metrics are most suitable for evaluating customer classification models that can provide a better result?

1.6 Research Significance

This study contributes to marketing analytics and customer segmentation by exploring the comparative effectiveness of predictive models in classifying customer value so that retail businesses can use similar analytical technique to segment their important customers to put more focus on them. Here, the findings can guide organizations in allocating resources towards the most profitable customer segments to improve personalization strategies and maximize customer lifetime value (CLV) using their value-added services towards those high valued customers.

2. Literature Review

2.1. Customer Value and Spending Behaviour

Customer value plays a pivotal role in the long-term success and profitability of any business, especially retail businesses whose operational growth depends on the retention and growth in the customer base. Fader, Hardie & Lee (2005) opined that high-value customers who tend to generate a disproportionately large portion of an organization's revenue, are critical for strategic planning so that businesses can put more emphasis on the service quality maximization for valued customers. Therefore, an early and accurate identification is essential for sustaining competitive advantage in increasingly data-driven markets. Ngai, Xiu & Chau (2009) argue that Customer Lifetime Value (CLV) remains a central metric in assessing customer worth, as it incorporates not only the total revenue generated by a customer over their relationship with the business but also factors in profitability, retention, and acquisition costs incurred to the business over the time.

According to Hosmer, Lemeshow & Sturdivant (2013), spending behaviour that comprises purchase frequency, transaction size, and recency of purchase serves as a foundational basis for assessing customer value as loyal customers are often assessed by such parameters. By analysing these behavioural metrics, businesses can estimate future purchasing trends and tailor marketing strategies

accordingly to attract more customers who have the potential to become loyal to the business. For instance, frequent buyers who consistently spend above the average transaction amount are typically segmented as high value who often receive personalized engagement efforts. Predictive analytics and historical purchase data analysis enables organizations to shift from reactive approaches to more proactive and data-informed customer targeting to get a competitive advantage over the other businesses in the market. As noted by Kumar & Reinartz (2016), integrating behavioural data with predictive models significantly enhances the ability to forecast future customer value with higher precision and it can lead a business towards successful implementation of their business growth planning in the market.

2.2. Traditional Customer Segmentation Methods

Wang & Yu (2017) found from research that traditional segmentation methods have long been the backbone of customer analytics for super shop chains and large retail shops around the globe. Techniques like RFM *(Recency, Frequency, Monetary)* analysis are simple to implement and provide intuitive customer groupings, so often businesses use such qualitative matters to evaluate customers manually. These methods are grounded in the assumption that recent and frequent purchasers with higher monetary contributions are more valuable to the business. Similarly, Blattberg, Getz & Thomas (2001) opined that k-means clustering has been extensively used for grouping customers into similar segments based on predefined variables from where the factual insights can be received from the business market. However, Chen et al. (2025) argued that such models are fundamentally limited in their ability to handle non-linear relationships and multivariate dependencies which are prevalent in real-world consumer dataset analysis to generate results. Moreover, these approaches often lack the sophistication to incorporate contextual or behavioural nuances about the customers or other external factors. For example, two customers might have similar RFM scores but very different product preferences or seasonal shopping patterns which causes such scenarios that traditional models cannot distinguish. In this case, Logistic Regression is a classical statistical method for binary classification which has also been utilized in segmentation tasks by researchers. According to Kailash, Kanwar & Sonia (2023), although it offers high interpretability and simplicity, Logistic Regression often assumes linearity between input features and the output, which can be a poor fit for complex datasets with high dimensionality and interactions among variables.

2.3. Machine Learning in Customer Analytics

With the introduction of big data in business environment and increased computational power provided by computers, machine learning (ML) techniques have revolutionized customer analytics for businesses. ML models excel in handling complex, non-linear, and high-dimensional data through uncovering patterns that traditional techniques often overlook. According to Gadgil, Gill & Abdelmoniem (2023), algorithms such as decision trees, random forests, support vector machines (SVM), and deep neural networks (DNNs) have demonstrated superior performance in predicting customer behaviours such as churn rates, repeat purchases, and lifetime value calculation for businesses to predict the retention of their customers and identification of their valued customers who can contribute more to their business.

Random Forest Model is an ensemble-based learning technique developed by Breiman (2001) which offers distinct advantages in terms of robustness and accuracy for analysing datasets. By combining multiple decision trees where each trained on a subset of the data, Random Forest Model can reduce variance and mitigate overfitting of values in the dataset. Such common issues in individual tree-based models can be easily mitigated through the usage of this model. Though Neural networks are more computationally intensive, they are capable of modelling highly complex patterns, particularly in unstructured data like customer reviews or social media interactions about a business. Despite their performance benefits, one challenge with advanced ML techniques is that the trade-off between accuracy and interpretability remains a critical concern for business stakeholders as it still requires actionable insights from the business.

2.4. Predictive Modelling for High-Value Customer Identification

According to Vallarino (2023), the use of predictive modelling for identification of high-value customers has gained traction in recent years among retail and small business ventures around the globe. Predictive models allow for more precise targeting by using historical behavioural data to classify customers based on their likelihood of high spending in the future if other factors remain fairly constant. Mena et al. (2019) found that ensemble models such as Random Forest Model outperformed traditional methods in precision and recall when identifying customers with high lifetime value in most cases. These findings underscore the potential of machine learning models in enabling more accurate and granular segmentation of customer behaviour and the valuation of customers in the market. Moreover, according to Sun & Wang (2024), predictive models can be retrained and updated dynamically as new customer data becomes available which ensures their continued relevance and accuracy with the existing analysis results. This real-time adaptability is particularly useful in fast-moving consumer goods (FMCG) and e-commerce sectors, where customer behaviour shifts rapidly in the business market. Beyond classification, Atay & Turanli (2024) opined that these models can also support in propensity scoring, upselling or cross-selling opportunities, and churn prediction of customers, all of which contribute to enhancing customer retention strategies and optimizing marketing budgets for a business in the competitive market.

2.5. Model Evaluation Metrics in Classification

Evaluating the performance of predictive models requires the use of appropriate classification metrics which will be used in analysing the existing information. Accuracy in datasets often may not be sufficient, especially when dealing with imbalanced datasets which is a frequent occurrence in customer analytics where high-value customers constitute a small fraction of the customer base. Precision, recall, and the F1-score provide more nuanced views about the dataset and the customer information in this case. According to Chamberlain et al. (2017), precision measures the proportion of true positives among predicted positives, which is crucial when false positives of misclassifying a low-value customer as high-value can incur high marketing costs for a business. Parallelly, recall focuses on identifying as many actual high-value customers as possible to ensure minimal valuable customers are missed. The F1-score

harmonizes these two metrics to provide a balanced assessment of the total findings from an analysis. Additionally, the ROC AUC *(Receiver Operating Characteristic - Area Under Curve)* metric is particularly useful in binary classification problems in dataset such as gender of Yes, No Scenario in the data fields. It evaluates how well the model distinguishes between classes across various threshold levels. Chen et al., (2024) stressed the importance of using a combination of these metrics for comprehensive model evaluation in skewed datasets where a single metric may not reflect overall performance about the whole scenario in the business market.

2.6. Challenges in Model Implementation

Rajagopal (2011) opined that despite varying capabilities, implementing machine learning models comes with notable challenges. One major issue is that complex models such as deep neural networks often function as "black boxes," making it difficult to understand how decisions are made as the interpretability of such models remain tough. This lack of transparency can hinder stakeholder trust and regulatory compliance in sectors with strict data governance protocols in the business environment. While models like Logistic Regression are less accurate, they are favoured in many practical applications for their straightforward interpretability in this case. Another challenge stated by Kailash & Kanwar (2023) is that ML models are highly sensitive to missing values, outliers, and inconsistencies in data, so, high-quality, clean, and pre-processed datasets is essential for accurate modelling in such cases. Feature selection also poses a challenge, as irrelevant or redundant features can degrade model performance. Feature engineering transforms raw data into meaningful input variables which requires domain expertise and iterative experimentation to understand and evaluate clearly. GhasemTanhaei et al. (2024) emphasize that even simpler models like Logistic Regression can benefit significantly from thoughtful feature selection and transformation which shows that data preparation remains a critical phase in any modelling process for ant type of business.

2.7. Literature Gap

While a substantial body of literature explores the use of machine learning for customer segmentation and behaviour prediction in sample competitive business environment, there is a notable lack of comparative studies that benchmark traditional models like Logistic Regression against more advanced ensemble methods such as Random Forests using real-world transactional datasets. Many existing studies either focus solely on a single model or use simulated data which limits the applicability of findings to real business environments in this case. Furthermore, few studies have examined how these models perform in the presence of class imbalance which is a common characteristic of datasets where high-value customers form a minority group in the business. Hence, this study aims to address these gaps by providing an empirical comparison of model performance on the same dataset, evaluating not only predictive accuracy but also practical implications for marketing decision-making by showing the results from the model implementation.

2.8. Conceptual Framework

The conceptual framework which has been used in this research involves the integration of structured customer transactional data into a predictive modelling pipeline. This framework includes data preprocessing such as handling missing values as well as feature selection, model training using Logistic Regression and Random Forest, and model evaluation using classification metrics in the process. Here, customers are classified as either high-value or non-high-value based on pre-defined thresholds of spending behaviour according to the transaction information from the dataset. The outputs from each model are compared using precision, recall, F1-score, and ROC AUC to determine which model provides superior classification performance. These results are then interpreted to guide marketing strategies, such as targeted promotions, personalized recommendations, and customer retention plans for a business to use in its strategic planning process. The framework is designed to be replicable and scalable which allows its adoption across various industry contexts for retail or FMCG businesses.

2.9. Research Gap

While numerous studies have examined the role of machine learning in customer analytics, very few have conducted comparative analyses that tests both traditional and modern machine learning models on current and relevant datasets for a specific type of business group. Moreover, there is a lack of research that focuses explicitly on the performance of these models in class-imbalanced environments, which is a critical limitation given the disproportionate nature of high-value customer distributions in a dynamic business environment. This study bridges this gap by conducting an empirical evaluation using a real-world dataset and a binary classification task in the dataset. By testing model performance across various evaluation metrics, this research contributes to the broader understanding of model suitability in practical business applications for retail businesses in the market.

2.10. Hypothesis Development

Based on the literature and theoretical framework, the following hypotheses are developed:

H1 (Alternative Hypothesis)

Random Forest shows higher performance metrics than Logistic Regression in predicting high-value customers based on their spending behaviour.

H0 (Null Hypothesis)

There is no statistically significant difference in the performance of Random Forest and Logistic Regression models in identifying high-value customers.

These hypotheses will be tested through empirical analysis with the goal of providing actionable insights for businesses which can help them in their customer segmentation and targeting strategies through predictive modelling.

3. Research Methodology

3.1. Introduction

This chapter discusses about the methodological approach adopted to analyse customer spending behaviour and identify high-value customers using machine learning models in this dissertation. This portion discusses the research philosophy, approach, strategy, design, sampling, data analysis methods, ethical considerations, and validity and reliability concerns related to this research to ensure the quality completion of the analytical process.

3.2. Research Philosophy

This research adopts a **Positivist Research Philosophy** which is grounded in the belief that reality is objective and can be observed, measured, and analysed through empirical methods. Positivism emphasizes the use of scientific and quantitative approaches to generate knowledge that favours observable and factual data over subjective interpretation. This process is well-suited to studies that rely on statistical analysis and measurable outcomes such as the evaluation of machine learning models in customer analytics in this dissertation. According to Vallarino (2023), Positivism aligns naturally with the nature of customer analytical study which involves the application of machine learning algorithms such as Random Forest and Logistic Regression to structure consumer transaction data. Mena et al. (2019) opined that these algorithms depend entirely on quantifiable input variables and yield numerical outputs such as probabilities, classification labels, and performance metrics. As the process is devoid of personal bias or human interpretation, further reinforcement of the objectivity required in a positivist framework to implement in the dissertation. By focusing on empirical validation of hypotheses and model performance using standardized evaluation metrics such as accuracy, precision, recall, F1-score, and ROC AUC, this research adheres closely to positivist principles to get the expected outcome with higher efficiency. The study seeks to identify consistent patterns and relationships within the dataset that can be generalized and potentially replicated in other business contexts so that the usage of the analytical process can provide usage in real life. Furthermore, the use of cross-validation and statistical analysis have strengthened the study's reliability and objectivity in the process.

3.3. Research Approach

This research has used **Deductive research approach** that starts with theoretical grounding and hypothesis formulation based on existing literature and prior empirical findings to get the final result. Here, hypotheses have been developed regarding the predictive performance of two machine learning models in classifying high-value customers from the analytical process. The deductive approach is appropriate here because the research seeks to test established theories related to customer segmentation and predictive modelling in a new context using real-world transactional data to get efficient results. Here, this approach supports empirical validation by analysing patterns in the data to either accept or reject these hypotheses. By aligning with a deductive framework, this research has maintained a structured and theory-driven path from problem identification to result interpretation using the approach.

3.4. Research Strategy

The research strategy utilized in this study is **Predictive Modelling** using **Quantitative data** analysis. According to Atay & Turanli (2024), predictive modelling is especially suited for tasks that involve forecasting future behaviour of customers or a target group based on historical data, such as identifying high-value customers based on past transactions in this regard. This research implements a data-driven strategy wherein Python language has been used to build, train, and evaluate classification models on consumer transaction datasets to derive results. The strategy here covers the entire machine learning pipeline from data exploration, preprocessing, feature selection, model training, evaluation to interpretation of results. Random Forest and Logistic Regression models were trained on the same dataset to ensure consistency and fairness in comparison in this case. The focus on quantitative strategy reinforced the objectivity of the research as numerical evidence forms the basis for all analyses and conclusions from the analytical process.

3.5. Research Method

This research employed a **Quantitative research method** as it was appropriate for numerical analysis and finding measurable outcomes from the analysis. Chamberlain et al. (2017) stated that Quantitative methods are valuable in machine learning research where model performance can be evaluated using standardized metrics and statistical comparisons from a certain dataset. This research utilized structured and tabular consumer transaction data that contains features such as purchase frequency, monetary value, recency, and customer demographics. Two supervised learning algorithms such as Random Forest Classifier and Logistic Regression have been implemented. The outputs were analysed using a variety of performance metrics including accuracy, precision, recall, F1-score, and ROC AUC that provided a multi-dimensional view of each model's effectiveness in classifying high-value customers in this dissertation. Besides, for ensuring quality of the outcomes, consistent data preprocessing, parameter tuning, and model evaluation protocols were applied in the analysis process. These steps ensured that results were not influenced by data inconsistencies or biases to ensure analytical integrity of the findings.

3.6. Research Design

Experimental and Comparative research design have been used here to empirically assess and compare the performance of two machine learning models in a controlled environment. This design facilitated a proper evaluation of the models under identical conditions to ensure the reliability and validity of the findings. Chen et al. (2024) stated that data preprocessing involves cleaning missing or inconsistent entries, encoding categorical variables, and normalizing numerical features where necessary so that the biases in the research can be mitigated from the process. The final dataset has been split into training and test sets using an 80/20 ratio to let the model understand the trend in the dataset and find a relation with the other external factors with the changes in the dataset. This allowed for the training of both models on the same subset of data while reserving an independent set for unbiased performance evaluation. To further enhance the robustness of the results, k-fold cross-validation was employed in analysis. This technique divides the dataset into *'k'* subsets and iteratively trains the model on *k-1* folds while validating it on the remaining fold in the analysis. This process is repeated *k* times, and the average

performance is taken to ensure consistency and prevent overfitting from the analysis. The use of both hold-out validation and cross-validation reinforces the methodological reliability of the comparative design in this case.

3.7. Sampling Technique

Purposive sampling technique was employed in the selection of relevant features from the dataset in this research. Unlike random sampling, purposive sampling is a non-probability sampling technique that involves the deliberate choice of variables that are believed to have significant influence on the outcome. In this case, features such as frequency of purchases, monetary spend, recency of last transaction, age, wealth segment, and gender were included based on their documented relevance in customer behaviour and marketing literature to find the correlations. As the dataset used here was pre-collected, anonymized, and ethically sourced, it has minimized the risk of selection and response bias from the analytical process. Since the dataset was drawn from public repositories, it helped to focus on modelling and analysis without concerns about data collection inconsistencies being collected in the real-world scenario. The selection of relevant, high-impact features helped increase the model's predictive power while maintaining interpretability in this context.

3.8. Research Limitations

This research acknowledges several limitations that may affect the interpretation and generalizability of the findings. Firstly, the analysis is limited to a single dataset of a particular business, which constrains the applicability of the results across different industries or market segments. A broader dataset encompassing multiple retail sectors or regions could improve the external validity of the research. Secondly, the class imbalance between high and low spenders posed a methodological challenge to evaluate. High-value customers typically represent a small portion of the population, which can skew model training and evaluation process. While performance metrics like ROC AUC and F1-score were used to mitigate this issue, future research should consider applying techniques such as SMOTE *(Synthetic Minority Over-sampling Technique)* or class weighting to address the imbalance more systematically and ensure more refined results. Thirdly, the interpretability vs. accuracy trade-off became evident in the comparison of models in this case, as Random Forest offered superior predictive performance, its black-box nature made it less interpretable than Logistic Regression. This limitation might impact its adoption in business contexts where transparency and decision accountability are crucial to derive results. Lastly, external demographic and behavioural variables such as customer location, seasonal trends, or product preferences were not included due to dataset constraints from the transactional dataset. Incorporating such features could potentially enhance the model's precision and contextual relevance for different businesses.

3.9. Ethical Considerations

Ethical integrity has always been a guiding principle throughout the research process. So, all data used in this study were fully anonymized so that no personal identifiers such as names, contact information, or financial details were accessible or stored that can affect any individual. The dataset was sourced from

public repositories, thus eliminating any potential breach of data privacy. In accordance with data protection standards, all data processing and storage were performed in secure environments, and no data was shared with third parties throughout the research. Moreover, the research strictly adhered to academic ethical guidelines to ensure confidentiality. Proper care has been taken to ensure that model outputs do not introduce bias or discrimination, especially regarding sensitive attributes like gender or socioeconomic status. Although such variables were included for analytical purposes, their impact on predictions was monitored and interpreted within an ethical framework to get expected outcome from the research.

3.10. Validity and Reliability

The validity and reliability of the research findings were carefully ensured so that the robustness and credibility of the study can be enforced. Here, internal validity was maintained by applying uniform preprocessing techniques and consistent model evaluation criteria. All features were treated equally across both models to ensure a fair comparative analysis for both models. Parameter tuning, model training, and testing were conducted under identical conditions so that threats to internal validity can be minimized. Moreover, External validity that refers to the generalizability of results may be limited due to the dataset's specific context. Since the data connect to a particular customer group or business model, results may not directly translate to different industries or customer types, but similar actions might provide a satisfactory result to a business in the similar industry.

Reliability has been ensured by implementing cross-validation and using replicable Python scripts for the entire modelling pipeline. The codebase used for data preprocessing, model training, and evaluation was well-documented and version-controlled that allows easy replication for other researchers or practitioners if they need it. The use of automated and consistent analytical tools ensures that similar results can be obtained under the same conditions, fulfilling the requirements of methodological reliability for other datasets.

4. Findings Analysis and Discussion

The application of data-driven predictive modelling has become integral in understanding customer behaviour within the retail industry. Leveraging the dataset from a retail company named **Sprocket Central Pty Ltd** which was available online at **Kaggle**, the goal of this research was to predict customer purchase behaviour, particularly identifying "*high spenders*" using demographic, transactional, and behavioural features. Using supervised learning techniques such as Random Forest and Logistic Regression, the analysis focused on classification accuracy, interpretability, and the underlying business insights derivable from the models that properly helped to understand the precision of this analysis process in retail businesses for getting important insights.

4.1. Dataset Overview

The dataset comprises three merged components which are transactions, demographic data, and address information. Key features that have been engineered here includes the following factors:

- *Recency*: Days since the last purchase.
- *Frequency*: Number of transactions.
- *Monetary*: Total value of purchases.
- *Age, Gender, Wealth Segment*: Demographic predictors.

Moreover, the binary target variable high spender was defined by segmenting customers based on whether their total spending exceeded the dataset's median value or not. In this way, it was transformed so that the machine can understand the changes in the data in different segments and the analysis can be done in a symmetric way.

4.2. Model Performance Metrics

In this dissertation, Random Forest Classifier and Logistic Regression models were trained and evaluated using test data to see if the preferred method outperforms the traditional Logistic Regression method or not. In this case, performance was evaluated using standard classification metrics such as accuracy, precision, recall, F1-score, and ROC AUC which have been discussed below:

4.2.1. Classification Report (Random Forest)

The classification report showed that Random Forest achieved strong precision and recall values from the training of the dataset behaviours. Few of the tools used here have been stated below:

1. *Precision (Class 1 - High Spender)*: The result indicates that the proportion of predicted high spenders were actually high spenders in that dataset.
2. *Recall (Class 1)*: This tool measures how many actual high spenders were correctly identified from the data fields.
3. *F1-Score*: This is the harmonic mean of precision and recall that provides a balanced measure for the analytical process.
4. *ROC AUC*: This analysis has a result value of 0.87 which demonstrated excellent separability between classes in this case.

4.2.2. Classification Report (Logistic Regression)

The Logistic Regression model is traditionally used for being more interpretable, showed slightly lower performance. The results show the below mentioned insights:

1. *Precision & Recall*: The result of this portion is lower than Random Forest that indicates more false positives and false negatives prediction by this model.
2. *ROC AUC*: At an approximate value of the result 0.81, it was still a good classifier, but less robust to interpret and evaluate complex and non-linear relationships among data fields.

4.3. Confusion Matrix Analysis after RF deployment

According to Gadgil, Gill & Abdelmoniem (2023), Confusion matrix provides a summary of the prediction results on a classification problem so that it can differentiate the number of true positives, true negatives, false positives, and false negatives from the analysis which helps in understanding the model's performance in detail. Below here, the confusion matrix analysis for the collected data has been shown below:

The confusion matrix stated above using Random Forest model revealed the following insights that has been stated below:

- High number of true positives are shown in the illustration, that means most high spenders were correctly identified by the model.
- A small number of false positives indicates that some low spenders were mistakenly identified as high spenders in the predicted dataset.

So, after careful judgement, a balanced accuracy across classes in this analysis process suggests the model was not biased towards any particular outcome throughout the process. Hence, this shows that effective modelling of customer behaviour with minimal misclassification is crucial for targeted marketing campaigns that is often preferred by retail businesses in competitive markets.

4.4. ROC Curve Insights

Hosmer, Lemeshow & Sturdivant (2013) stated that the ROC curve plots the true positive rate against the false positive rate at various threshold settings so that AUC can provide a single measure of the model's ability to distinguish between classes. Here, in analysing results, higher AUC indicates better model performance. The ROC curve analysis between random forest and Logistic regression based on the collected dataset has been shown below:

The above graph information has been discussed below:

- Random Forest Curve displayed a convex shape towards the top-left which indicates high TPR with a low FPR.
- Logistic Regression Curve here can be seen as slightly flatter which confirms less discriminatory power compared to the ensemble model.

So, these plots visually confirm that the Random Forest model performs superiorly for marketing scenarios where identifying actual high spenders is crucial to minimizing advertising costs and ensure value addition services to turn them into loyal ones.

4.5. Feature Importance Interpretation

Understanding which features contribute most to the predictive performance of a model is vital for both interpretability and business decision-making for every business. In this research, the Random Forest

model has been used to understand feature importance using the built-in feature importance scores based on mean decrease in impurity. The analysis provided several critical insights regarding which variables most strongly influenced the model's ability to classify high-value customers based on the correlations among the variables and their changes. Below here the graph has been shown:

Monetary

As anticipated, the 'Monetary' variable which represents the historical spending amount of each customer has been found as the most influential predictor in identifying high-value customers. This aligns closely with foundational marketing theories that suggest that past monetary contributions are a strong indicator of future value for a customer. High spenders in the past are likely to continue spending at similar or increasing rates unless there is a shift in circumstances which makes this feature a reliable proxy for customer potential to become loyal ones.

Frequency

The second most important variable was 'Frequency,' which refers to the number of transactions a customer has made within a specified time frame to the business. Customers who shop more frequently demonstrate a higher level of engagement with the brand that indicates not just interest but potential loyalty from the target groups. Frequent transactions often signal sustained value over time from customers that ensures a smooth growth in the profitability of the business. Wang & Yu (2017) have found that frequency also interacts well with recency to create a composite view of behavioural momentum among the customers of a business.

Age

The customer's age has been identified as a moderately strong predictor as positive correlation was observed between age and high spending that suggests that older customers may have greater purchasing power, possibly due to higher disposable income, established careers, or stable financial situations but age alone does not determine customer value, when contextualized with other factors.

Recency

The recency variable measured how long ago the customer last made a purchase. In this case, the analysis shows that Recency exhibited a negative correlation with high-value classification which indicates that customers who purchased recently were more likely to be classified as high spenders. This supports marketing practices that prioritize engaging recently active customers who are more likely to respond to offers and campaigns.

Wealth Segment

This categorical variable stratified customer based on their socioeconomic status to look into broader consumer behaviour in the target segment. Customers who have been classified under higher wealth

segments were more likely to be high-value spenders. Here, this variable added granularity to the model by highlighting how spending patterns can differ across income brackets that confirms the long-held assumptions in consumer economics in the business market.

Gender

Gender in this analysis has a measurable influence on model performance. Although its impact was relatively small, the inclusion of gender enabled the model to capture subtle patterns that may relate to shopping preferences or product category affinity which could differ across male and female customers in certain retail settings in the business market.

4.6. Business Implications

The successful identification of high-value customers through predictive modelling offers wide-reaching strategic and operational benefits for organizations in competitive retail or service environments. The insights derived from such models extend well beyond academic interest and serve as vital tools for improving business performance, profitability, and customer experience. The primary benefit lies in the ability to segment customers more precisely to drive sales in the business. According to Kumar & Reinartz (2016), by identifying high-value individuals, organizations can tailor marketing messages, offers, and incentives specifically to the target group. Personalized promotions, loyalty perks, and exclusive access to new products can be directed toward customers which ensures optimal allocation of marketing budgets. This targeted approach enhances both engagement and conversion rates, reducing customer acquisition costs. For example, in this dissertation, the initial transaction dataset had inputs of more than 20,000 rows from 4000 + customers. From those customers, based on their purchase habits, frequency and other factors, the model has selected 600 + customers who are valuable for targeting to turn them into loyal ones. The output result is shown below:

Another key implication is the potential to develop data-driven retention programs by the businesses as Customers exhibits declining engagement sometimes which can be flagged as at-risk. This enables timely intervention through reactivation campaigns, personalized outreach, or loyalty program incentives from the outcome from the analysis. According to Kailash & Kanwar (2023), retaining an existing customer is significantly more cost-effective than acquiring a new one that makes this a crucial strategic advantage for every business. Moreover, analysis of high-value customer behaviour provides insights that can feed directly into product design, development, and pricing strategies for a business. For instance, older customers might spend more could prompt product lines or packaging specifically tailored to that demographic so that they can be positively assisted to purchase. Similarly, gender-related insights may influence assortments or promotional tactics by ensuring product-market fit which increases customer satisfaction in this case.

Though this research focused on classifying high spenders, the models developed can be extended to predict Customer Lifetime Value by integrating time-variant and contextual data with predictive modelling to develop long-term revenue projections from customers. These projections are instrumental

in long-range planning, capital allocation, and investor reporting for a business to address their current opportunities in the market. Beyond marketing, departments such as sales, customer service, and supply chain can benefit from these insights as through identification of demand patterns tied to high-value customers can improve inventory forecasting and logistics planning to reduce costs in the business operations.

4.7. Comparison of Models

The empirical comparison between Random Forest and Logistic Regression models offers valuable insights into their respective strengths and limitations for high-value customer prediction process. The performance metrics were used to evaluate and contrast both models on the same dataset below here:

Metric	Random Forest	Logistic Regression
Accuracy	High	Moderate
Precision (Class 1)	High	Moderate
Recall (Class 1)	High	Slightly lower
ROC AUC	0.87	0.81
Interpretability	Moderate	High

According to the above table, the Random Forest model clearly outperformed Logistic Regression in most predictive metrics. Its high ROC AUC score of 0.87 indicates strong discriminative ability across classification thresholds, while its superior precision and recall reflect its effectiveness at correctly identifying high-value customers. This is very important in marketing contexts where misclassifying low-value customers as high-value can lead to inefficient spending by retail businesses. On the contrary, Logistic Regression's interpretability remains a strong advantage in regulated industries or when model transparency is essential. Also, business stakeholders may find it easier to understand and trust the outputs of a logistic model when decisions must be auditable and defensible. Therefore, while Random Forest is more suitable for maximizing predictive performance, Logistic Regression may still be preferred in scenarios prioritizing explainability over complexity based on the context of different businesses in the competitive landscape. Hence, hybrid approaches or model stacking could be considered to balance performance and interpretability if needed for any business venture.

4.8. Methodological Limitations

Although the findings from this study are encouraging, there are different methodological limitations that needs to be acknowledged as they may impact the generalizability and applicability of the results in different businesses. Firstly, the dataset used in this case showed class imbalance which is a common issue in customer analytics where high-value customers represent a minority. In such cases, overall accuracy becomes a misleading metric, as models can achieve high accuracy by simply predicting the majority class from a dataset. Though metrics like precision, recall, and ROC AUC were used to

counteract this, future studies should incorporate techniques like SMOTE, ADASYN, or under-sampling to ensure balanced training to the selected models (Chamberlain et al., 2017). Secondly, customer preferences and behaviours evolve over time due to economic shifts, seasonality, and changing market conditions in the business market. The models used here were trained on static historical data, which may limit their validity over time. So, through incorporating temporal modelling techniques such as time-aware neural networks or Boost with lag features could help improve long-term accuracy and adaptability with the changes in dataset.

Though key transactional and demographic variables were included, several potentially predictive features were omitted such as location data, device usage patterns, loyalty program interactions, and product-level preferences. The exclusion of such variables may limit the model's holistic understanding of customer behaviour to purchase products from the selected business. Moreover, although feature importance was analysed, advanced explainability tools like SHAP *(SHapley Additive exPlanations)* or LIME *(Local Interpretable Model-agnostic Explanations)* were not applied to reduce complexity in the analytical process. But these tools could provide richer, instance-level explanations which might enhance trust in black-box models and aiding stakeholders in decision-making with more technical indications (Mena et al., 2019).

5. Recommendations and Future Research

Based on the insights and limitations identified, several recommendations for future research have been suggested below to further enhance the predictive modelling framework and its real-world applicability:

- Future models should incorporate additional behavioural and contextual features such as average basket size, preferred shopping channels, coupon usage, and browsing data to get more dynamic information about their customers. Incorporating dynamic features like time of purchase, weekend vs weekday behaviour, and event-based promotions can significantly enrich model performance while doing the analysis on models.
- Though Random Forest provided high accuracy, gradient-boosting techniques such as XGBoost, LightGBM, and CatBoost should be explored as these models often outperform Random Forest in structured data tasks and offer improved performance in handling class imbalance, missing values, and regularization of the data fields in the dataset (Rajagopal, 2011).
- SHAP or LIME can bridge the gap between predictive performance and interpretability in further research as these tools can provide both global and local explanations that provides stakeholders with information to understand why a specific customer was classified as high value (Gadgil, Gill & Abdelmoniem, 2023).
- Combining predictive modelling with unsupervised learning techniques like K-means or DBSCAN can identify behavioural clusters within high-value customers for getting a more specific and precise selection of high value customers. This layered approach supports more nuanced targeting

strategies where marketing campaigns can be tailored not only by value but by shared preferences or habits within segments.

6. Conclusion

This research tried to explore the effectiveness of machine learning techniques in identifying high-value customers by analysing transactional and demographic data. Through a comparative analysis of Random Forest and Logistic Regression models, the research aimed to address existing gaps in literature regarding model performance, interpretability, and applicability in real-world with imbalanced datasets which are mostly common among real world data. However, the findings confirm that Random Forest offers superior predictive accuracy compared to traditional statistical approaches like Logistic Regression model. Key features such as historical monetary value, transaction frequency, customer age, and recency were found to be strong indicators of high customer value by analysing them with RF model. These results align with established retail marketing theories which supports the importance of past behaviour and demographic traits in forecasting future customer worth for a business. The research also highlights several practical implications for businesses as accurate prediction of high-value customers enables more precise marketing segmentation, improved customer retention strategies, and informed product development decisions by a business. Furthermore, the capability to extend these models toward Customer Lifetime Value (CLV) prediction opens new avenues for long-term strategic planning for retail businesses. Moreover, the existing limitation regarding class imbalance, static datasets, and limited feature scope for this research has also been pointed out so that future work can aim to enhance the robustness of predictive models through time-aware techniques, richer feature sets, and the integration of explainability tools like SHAP or LIME to get more efficient results. However, this research tried to discuss the transformative potential of machine learning in customer analytics in a dynamic business environment. By moving beyond traditional segmentation methods, businesses can leverage predictive models not only to identify their most valuable customers but also to drive intelligent, data-informed decision-making across marketing, sales, and strategy functions in competitive business environment.

Ethics, Consent to Participate, and Consent to Publish declarations: not applicable. This study used a fully anonymised, publicly available dataset (Kaggle) containing no personal identifiers; therefore, ethical approval and informed consent were not required.

Declarations

Ethics, Consent to Participate, and Consent to Publish declarations: not applicable. This study used a fully anonymised, publicly available dataset (Kaggle) containing no personal identifiers; therefore, ethical approval and informed consent were not required.

Ethics, Consent to Participate, and Consent to Publish

Clinical trial number: not applicable.

Clinical Trial Number

This research received no external funding. The authors declare that no funding was provided for this study.

Data Availability

The data that support the findings of this study are available from KPMG via the KPMG Data Analytics Consulting Virtual Internship dataset

(https://www.kaggle.com/datasets/sudiptog81/kpmg-virtual-internship), but restrictions apply to the availability of these data, which were used under licence for the current study, and so are not publicly available. Data are, however, available from the authors upon reasonable request and with permission of KPMG.

Author Contribution

Raihan Sharif: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & EditingSalmi Mohd Isa: Supervision, Conceptualization, Writing – Review & EditingAshik Chowdhury: Supervision, Methodology, Writing – Review & EditingMohammad Nasir Hossain: Supervision, Validation, Writing – Review & Editing

References

1. Atay MT, Turanli M. Analysis of customer churn prediction using logistic regression, k–nearest neighbors, decision tree and random forest algorithms. Advances and Applications in Statistics; 2024. pp. 1–15.
2. Blattberg RC, Getz G, Thomas JS. Customer equity management. J Mark. 2001;65(1):1–17.
3. Breiman L. Random Forests Mach Learn. 2001;45(1):5–32.
4. Chamberlain BP et al. 2017. Customer lifetime value prediction using embeddings. In: Proc. KDD Workshop, pp.1–10.
5. Chen, et al. Results in Control and Optimization: Predictive Analytics in Customer Behavior—superior RF & LR models. Results Control Optim. 2024;17:100462.
6. Chen, et al. Predictive analytics in customer behavior in Thai e–commerce: evaluating Random Forest and Gradient Boosting. J Global Rev Value R&D. 2025;14(1):1–20.
7. Fader PS, Hardie BG, Lee KL. RFM and CLV: Using iso–value curves for customer base analysis. J Mark Res. 2005;42(4):415–30.

8. Gadgil G, Abdelmoniem. 2023. A Meta–learning based Stacked Regression Approach for Customer Lifetime Value Prediction. arXiv preprint arXiv:2308.08502.
9. GhasemTanhaei H et al. 2024. Predicting customer behavior in Thai e–commerce: RF & GB models outperform logistic. JGRV, 14(1).
10. Harshit Kailash K, Sonia. 2023. Machine Learning Algorithms for Predicting Customers' Lifetime Value: A Systematic Evaluation. Proc. Conference Paper, May; available via ResearchGate.
11. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd ed. Hoboken: Wiley; 2013.
12. Kailash K, Kanwar S. 2023. Machine Learning Algorithms for Predicting Customers' Lifetime Value. Conference Paper.
13. KPMG. (2020). KPMG Data Analytics Consulting Virtual Internship: Sprocket Central Pty Ltd transactions dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/sudiptog81/kpmg-virtual-internship
14. Kumar V, Reinartz W. Customer Relationship Management: Concept, Strategy, and Tools. 3rd ed. Springer; 2016.
15. Mena et al. 2019. Churn Prediction with Sequential Data and Deep Neural Networks: A Comparative Analysis. arXiv preprint arXiv:1909.11114.
16. Ngai EWT, Xiu L, Chau DCK. Application of data mining techniques in customer relationship management: a literature review and classification. Expert Syst Appl. 2009;36(2):2592–602.
17. Rajagopal S. 2011. Customer Data Clustering using Data Mining Technique. arXiv preprint arXiv:1112.2663.
18. Sun P, Wang Y. The impact of online purchase behaviour on customer lifetime value. Journal of Marketing Analytics; 2024.
19. Vallarino D. 2023. Buy when? Survival machine learning model comparison for purchase timing. arXiv preprint arXiv:230814343.
20. Wang Y, Yu C. Machine learning in marketing: literature review and classification. Int J Inf Manag. 2017;37(2):87–101.