National College of Ireland

Comparative performance of RF and GBM for short-term customer segmentation forecasting

MSc Research Project
Cloud Computing

Thomas Jose
Student ID: x22146962

School of Computing
National College of Ireland

Supervisor: Dr Giovani Estrada

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Thomas Jose
Student ID:	x22146962
Programme:	MSc in Cloud Computing
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr Giovani Estrada
Submission Due Date:	25/04/2024
Project Title:	Comparative performance of RF and GBM for short-term customer segmentation forecasting
Word Count:	7566
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Thomas Jose
Date:	26th May 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Comparative performance of RF and GBM for short-term customer segmentation forecasting

Thomas Jose
x22146962

Abstract

The research explores the use of Recency, Frequency, and Monetary (RFM) analysis for customer segmentation. Companies often use segmentation techniques to generate insights on purchase behavior to quantitatively rank and group customers for targeted marketing campaigns. A typical question analysts face is how much data they need to perform those analyses and the confidence of predictions. However, little is known about what month is the easiest one to classify. Or, what is the best month in which the largest share of target customers is found? Here we present a detailed analysis of 1-month customer segmentation forecasting using two well-known machine learning techniques. We critically evaluate the classification accuracy on three segments, for "good", "medium" and "bad" customers. Based on the findings, we also evaluate the need for automated model selection and hyper-parameter optimization of the customer segmentation models. While many papers would go straight into the optimisation part, we want to verify whether such a range of tools is actually needed for improved classification accuracy. Based on the literature review, Random Forest and GBM classifiers were singled out as the top techniques for this classification task. It is however not known which one will deliver the best classification accuracy. As part of the research, we will show that plain RF and GBM are unsuitable for the task, as the distribution of "good" customers is uneven. To avoid this class imbalance, we used stratified classification.

The findings of the study point toward that, although April is the best month for prediction with RF, discussions have highlighted issues such as the classification of a few "good" customers and small sample sizes, and the need to tune the RFM score function or the "good" customers threshold. Besides, GBM outperforms RF by far, especially in those months that have a smaller number of "good" customers. GBM and RF give 98.5% and 91% classification accuracy, respectively with stratified classification using 10-fold cross-validation. Models are good enough, well over 90% accurate, and hence there is no need for further boosting of hyper-parameters. The results and research methodology are expected to provide valuable insights for analysts planning to do customer segmentation and forecasting of customer behaviour.

1 Introduction

1.1 Problem Background

Every company prioritizes value creation because it enhances the client experience (Sebald and Jacob; 2020). Several studies have proved that producing value requires knowing

customer needs at every stage. Customer intelligence can be defined as the process of using algorithms and tools to understand customers from a huge dataset of customer activity (Dam et al.; 2021). Customer classification and strategic data-driven decisions require acceptable approaches in the big data age. Classifying consumers properly will help strategic decisions in a competitive setting that prioritizes high-value clients (Chiliya et al.; 2009). Hughes introduced RFM in 1994 after Stone and Bob enhanced it in 1989. It calculates subscriber life value and loyalty using recency (R), frequency (F), and monetary (M) components (Gustriansyah et al.; 2020). To address their demands as they arise, customer classification studies customer behavior over time. This requires segmentation using behavioral data, which is readily available, constantly changing, and based on early purchases. Recency, Frequency, and Monetary (RFM) analysis is a popular way to evaluate clients' purchase habits (Christy et al.; 2021). A behavioral-based data mining method that creates customer profiles based on recency, frequency, and money was presented in (Tavakoli et al.; 2018). This method is used before client segmentation. One of the best marketing methods is categorizing customers by purchasing habits and traits. The study segments customers using RFM modeling. RFM Modelling helps companies assess consumer behaviour based on recency, frequency, and monetary value. This segmentation method divides clients into homogeneous groups so they may interact with different groups utilizing focused marketing. Quantitative analytical models like the RFM model are vital in CRM. The RFM model uses recency (R), frequency (F), and monetary value (M) to describe customer importance and type. To find potential customers, companies use the RFM model and past data to analyse client sales and buying behaviour (Huang et al.; 2020). Additionally, clients rate product or service quality from 1 to 5, with 1 being the highest and 5 being the lowest. In the same way, customers can be rated from 1 to 5 using percentiles of R, F, and M components. E-commerce sites evaluate products based on customer feedback, revealing product performance. However, these reviews will enhance the seller's service. For historical data, this research predicts the customer's future purchase review score into three cluster groups (good, medium, bad) with each cluster representing the type of customer based on the dynamics between R, F and M values.

1.2 Motivation

Customers' actions depend on demographics, social trends, company marketing methods, and government policies. Thus, a segmentation model that adapts to new behavior patterns is quite valuable for managers. Businesses usually need to manually develop a segmentation model, which might lead to customer segmentation issues. The managers should mix segments at the end of Quantile Analysis models to create meaningful segments. If there is a struggle to combine Recency, Frequency, and Monetary correctly and identify the optimal attributes for clustering-based client segmentation, a static segmentation parameter may be needed for comparison. For instance, fixed segments can help to compare segment properties over time to determine success. Existing RFM models solely cluster consumers with a set threshold, ignoring data form and user behavior similarities. Therefore, the results cannot provide high-quality user segmentation. Recency, Frequency, and Monetary are treated as independent characteristics in RFM models, and their relationship is not considered relevant. Thus, a relationship-focused segmentation methodology helps managers understand their segments. A typical task for the analyst is the prediction of demand and/or sales. Depending on the corporate strategy, this forecast

can be done for short-term (1-month or 3-month period) or longer-term (6-month and years ahead). We take the shortest possible forecast (immediate, 1-month ahead) due to limitations in the dataset, but the methodology could be applicable to longer ones.

1.3 Research Questions

In what manner can RFM analysis and ML algorithms like RF and GBM be evaluated to identify the best month for customer behavior forecasting, while also resolving the class problems that may arise in predicting 'good' customers throughout the year?

1.4 Research Objectives

To address the research question, four research objectives have been framed to add significant value to the research as follows:

1. Create different dataset splits as to be able to identify the best month to forecast customer behavior using a Random Forest (RF) classifier. It will be based on 2 months of training data and 1 month for forecasting, using a sliding window over the customer purchase dataset.

2. Compare the results of our baseline classifier (RF) with GBM classifier.

3. Addresses the low numbers of "good" customers in certain months and its impact on forecasting customer behavior by means of stratified classification using a 10-fold cross-validation approach. A comparative analysis of RF and GBM stratified classifiers is performed to find the best among them.

4. Provides insights into the research findings based on the classification accuracy as regards to the need for AutoML or hyperparameter optimization. The research implications are discussed in detail along with its limitations.

1.5 Thesis Structure

Section 1 presents the research projects, motivation, research question and objectives of the research. Section 2 details all the existing literature and their research gap that this research tries to cover. Section 3 provides the methodology for customer segmentation using KDD process. Section 4 details the design specifications of the research problem. Section 5 provides the implementation of customer segmentation models using RF and GBM organized as three case studies. Section 6 showcases the evaluation of the case studies and analysing the classifier performance for good, bad, and medium customers. Section 7 concludes the research with future scope.

2 Related Work

Businesses must make decisions targeting customers using empirical evidence like finding patterns from available data. Many machine learning (ML) algorithms exist in literature that can utilize this data to create models capable of segmenting customers based on their purchase behavior. Many applications use data mining, RFM analysis, customer

value matrix, and customer lifetime value (CLV) for customer segmentation. This literature review primarily focused on RFM analysis and clustering algorithms used by other researchers over the years in various domain areas of research including customer segmentation.

2.1 RFM Analysis on Customer Behavior

The authors have reviewed data mining strategies for consumer segmentation using RFM model in (Ernawati et al.; 2021). They examined 44 Scopus, Web of Science, and emerald papers published from 2015 to 2020 that satisfied their inclusion criteria. RFM was the most employed model for consumer behavior analysis and customer segmentation, according to their review. About 57% of research work used the basic RFM model, while others redefined or added variables to suit their unique applications. Most data mining studies used clustering approaches (88%) like K-Means, followed by visualization. In conclusion, the researchers comprehensively examined data mining approaches in RFM-based consumer segmentation across domains, found patterns, and offered an integrated framework for future research and implementation. Two consumer segmentation models for a Turkish sport retailing company was presented in (Dogan et al.; 2018) utilizing RFM analysis and clustering approaches to overcome the inadequateness of the company's customer spending based segmentation model. They recommended adding behavioral indicators like recency, and frequency of the purchase to understand the customer segments better. The study utilized 700,032 loyalty card customer records to compute the RFM scores. The two-step clustering methodology categorized customers into Bronze* (low R, F, M), Gold* (high R, low F, M), and Premium* customers (high R, F, M). This also identified 60% discrepancy in the existing segmentation model indicating many customers had been misclassified. This new model was found to be more trustworthy and actionable compared to the company's spending-based customer segmentation model.

Using RFM analysis to analyze customer purchasing records, (Chen et al.; 2009) presented a new sequential pattern mining (SPM) algorithm, RFM-Apriori. A pattern segmentation approach was presented that used mined RFM sequential patterns to provide management decision-makers insights into customer purchasing behavior. This approach outperformed the generalized sequential pattern (GSP) algorithm in identifying a reduced set of high-value patterns while retaining runtime speed. The scalability analysis revealed that both algorithms' runtime and number of patterns scale gradually with the number of customers, but exponentially increase with the average number of transactions and products bought in each transaction. To estimate the sales of many products using pharmacy sales data, an RFM and Fuzzy Analytic Hierarchy Process (FAHP) based sales prediction model was suggested by (Gustriansyah et al.; 2017). A case study was presented involving 6,877 product items, 127,047 sales orders, and 399,738 sales order information over an year. The model considered expert opinions, data preprocessing, criteria scaling and scoring through mean absolute percentage error (MAPE) for prediction accuracy assessment resulting in high accuracy with an average MAPE score of 3.22% for sales quantity forecast. According to the authors, this model could improve the efficiency of pharmacy inventory management process. For future studies, it was suggested that AI-based predictive models should be compared to determine the most accurate model for revenue estimation, specifically when dealing with complex data patterns as that presented in the case study.

Online customer segmentation has been addressed by a multi-layer RFM (MLRFM)

model proposed by (Handojo et al.; 2023) to overcome the limitations of RFM in fairly differentiating between old and new customers, especially that, existing customers have churned or stopped purchasing, while new customers have just started. This technique layered time frames and evaluated RFM values. The authors utilized numerical examples to demonstrate its usefulness for online retailers with dynamic consumer base. This research was compared with other existing RFM methods revealing that it ranks new high-activity customers above less active ones, thus giving more accurate priorities for current users and recommended further investigation into its adoption and evaluation in online transactions such as health care services and transportation systems. In 2018, an upgraded RFM model known as RF+M (Tavakoli et al.; 2018) for customer segmentation that defined recency segments in a semi-dynamic way by separating the recency factor from frequency and monetary variables to solve the problem of RFM analysis. The customers were then clustered through frequency, monetary or their linear combination using K-means clustering. They tested their R+FM model using data from Digikala, which is the largest e-commerce company in the Middle East. For active customers, three 'recency' groups were identified by the researchers (Active, Lapsing, Lapsed) and four 'frequency-monetary'segments (High Value; Medium Value with High Monetary; Medium Value with High Frequency; Low). Based on the increased purchase rates and average order values compared to previous years, the R+FM model was more effective than previous RFM campaigns used by Digikala.

The research conducted by (Dursun and Caber; 2016) utilized stratified sampling with RFM analysis to profile successful hotel customers. Out of 5939 total clients, they selected only 369 hotel guests who were mostly couples from Russia and Germany aged between 35-44 years old with children. Eight customer clusters were identified through RFM analysis where most of them belonged to 'Lost Customers' category (36.0%) having low scores of R/F/M while second highest cluster was represented by 'New Customers' (25.7%) followed by "High Potential Customers" at rank three accounting for 21.9%. Based on the size of various segments as well as their market potentials, the managers were advised to design service plans around the findings. The authors also suggested collecting additional customer information from different service points so that better profiles can be created about them. An RFM based consumer lifetime value (CLV) model for LWC Company, a paint distributor was presented in (Monalisa et al.; 2019). The researchers created transaction-based client portfolios to help the organization differentiate its CRM strategies. Fuzzy C-means clustering algorithm was used to cluster consumers on the basis of RFM model attributes. The results indicate that three types of clients were identified namely superstar, typical and dormant segments at various levels within LWC Company. Moreover they recommended that consumer management should be done by portfolio type at LWC Company. The findings emphasized the importance of using customer transaction data to understand behavior and develop tailored CRM strategies.

2.2 Machine Learning Models for Customer Segmentation

Customer segmentation analysis with machine learning classifiers has been a subject of great interest because it reveals hidden structures in customer data that allow companies to personalize marketing campaigns better. This section will review various research on the employment of ML classifiers in customer segmentation with special emphasis on the methods and results.

An empirical statistical analysis and discussion of the predictive abilities of selected

CLV models that could be used in online shopping within e-commerce business domain was presented in (Jasek et al.; 2018). The predictive abilities of the Extended Pareto model, Status Quo model, and Markov chain model for LTV were analyzed by training them on six online retail datasets that generate millions of euros in annual revenues. The evaluation metrics performance of the extended Pareto model was excellent and stable than the other two models used in this analysis, specifically for non-contractual online shopping cases. This implies that the extended Pareto model is most ideal for CLV prediction in real-time market scenarios. The authors, (Ansari and Ghalamkari; 2014) recommended using the RFM model and LTV to categorize mobile sales website customers. After calculating and weighting RFM values using Shannon entropy, the authors clustered customers with cluster CLVs into four primary parts using self-organizing maps (SOM). According to the survey, Cluster 2 has the most valuable clients, followed by Clusters 1, 4, and 3. They recommended using larger datasets over longer time-periods for better findings and deeper customer behavioral insights and sought to identify critical customers, and improve CRM.

The authors of (Rungruang et al.; 2024) presented an approach to customer segmentation using the RFM model and formal concept analysis (FCA). FCA has been utilized for creating knowledge representation that takes into account both implicit as well as explicit information. The explicit information was represented in the hierarchical structural model, while embedding its implicational characteristics into hidden knowledge. They then compared their proposed method with K-means and hierarchical clustering on online retail II dataset from UCI Machine Learning Repository. The results indicated that this method gives marketers enough relevant knowledge about where different types of customers are in relation to one another thus enabling them come up with practical marketing strategies for businesses in the real world by simultaneously segmenting people based on their relationship with each other as well as themselves. A research that relied on business intelligence to identify retail customers by evaluating sales history and purchase behavior was proposed by (Anitha and Patil; 2022). The RFM model and K-Means clustering technique were used to segment datasets and validating the dataset clusters with the Silhouette Coefficient. The study used a real-time transactional retail dataset over a defined business transaction duration. The dataset values and parameters arranged the regional client buying trends and behavior. The curated and organized data from this research study increased the company sales and profit and provided intelligent insights into customer purchasing behavior and related patterns.

An RFM model and K-means clustering technique to segment B2B insurance customers was presented in (Kumar and Philip; 2022). They adapted the RFM model for insurance, employing tenure for recency, number of policies for frequency, and profit margin multiplied by annual premium for monetary value. The authors preprocessed 127,037 consumers to extract RFM-relevant features, used elbow and silhouette modeling to find the ideal cluster number. Clusters were interpreted using RFM as: Cluster 0 - with loyal clients, varied portfolios, and low-profit margins for new products, Cluster 1 for premium product access due to its most loyal and valuable customers, Cluster 2 with loyal customers with poor profit margins, Cluster 3 with high-profit new entrants who may be targeted for new product marketing. The authors recommended organizations adjust the model by varying RFM parameters based on their targets and suggested more micro-segmentation analysis using more data. A new consumer segmentation method employing RFM data using K-means and Fuzzy C-means clustering algorithms was developed by (Christy et al.; 2021). The authors presented Repetitive Median based K-means (RM K-means) to reduce

iterations and improve cluster compactness compared to existing clustering techniques on a one-year transactional online retail dataset. The authors clustered clients by RFM score using K-means, Fuzzy C-means, and RM K-means. The algorithms' iterations, execution time, and average silhouette width were assessed, and found that the execution time was lower for the RM K-means algorithm than the other two approaches, suggesting that RM K-means algorithm outperformed standard K-means and Fuzzy C-means clustering algorithms. The study in (Frasquet et al.; 2021) was about consumer purchase behavior subsequent to a Spanish grocery store establishing an online store, whereby the firm used customer records during 18 months before and after the introduction of the online purchase option. The researchers segmented customers by means of latent class analysis (LCA) according to their buying behavior with an improved RFM model which took into account pre-channel transactions and was applied on 1151 customers who had done shopping on the new online channels. The authors identified seven significant groups among customers that showed different behaviors both offline and online. While three clusters (47%) were more or less open towards digitalization, two categories (35%) gave up on it after some time. Another category (13%) showed considerable shift in shopping places – moving mainly from brick-and-mortar stores into virtual ones and finally, one group (5%) churned soon after its adoption. In addition, the membership segments were found to have been significantly impacted by the volume of transactions prior to online activities in brick-and-mortar stores. This implies that those who spent a lot before online store presence also showed signs of engaging with both online and offline channels later on. Even people who already comprised the highest-spending groups increased their expenditure levels post-online channel establishment.

The consumer behavior in Iran's traditional and computerized banking channels was explored by (Hosseini et al.; 2022). A revolutionary two-dimensional approach to consumer transactions employing the RFM model was introduced using expectancy-value theory. One million transaction records from 85,000 customers of Iran's largest private bank, XYZ, were collected for one year. CRISP-DM was used to extract knowledge from the data, and K-Means clustering was used to segment clients by RFM in traditional and electronic channels. Shannon's entropy was used to determine RFM parameter weights. The researchers then computed each cluster's CLV utilizing Lawshe criteria. They combined the channel clusters to find 16 new two-dimensional client groupings. Customers who used both channels and had significant account balances had the highest CLV, indicating they were the bank's most valued and loyal. The study found that customers who used conventional and internet banking services generated higher revenue for the bank. A research work, detailed in (Shokouhyar et al.; 2020) investigated the relationship between customer satisfaction and after-sale service quality of cars. Researchers employed different components of service quality to evaluate how this affects overall customer satisfaction using Service Quality (SERVQUAL) model. They also used Fuzzy Kano Model which classified twenty-one quality factors and RFM clustering to handle customer attitude variations and measure the impact of post-purchase services on customer satisfaction in a heterogeneous group. They discovered that each cluster had its own peculiar preferences with regard to qualities. For example, group 2 considered "general attitude and behavior of technician" as an important factor, while group 1 was indifferent on its opinion. Additionally, data showed that some factors such as availability of service personnel were critical in avoiding dissatisfaction among customers, whereas proximity to the service centre greatly influenced satisfaction for some cluster groups.

2.3 Summary

The existing literature discussed in this section has been good at discovering elaborate patterns in client data and calling for individualized marketing efforts. However, from deeper analysis, it was learnt that the main research gap evident from this research is that not much thought has been laid to analyze customer behavior in the case of small datasets where the historical data pertaining to the customer behaviour may not be available. Hence, this research aims to address that research gap identified in the two closest works (Huang et al.; 2020) and (Tavakoli et al.; 2018) on which this research is based on, by developing classification models to be trained on parts of data instead of the whole dataset to determine if a small volume of data can be used effectively to make accurate predictions of the customer segments without compromising on the model evaluation performance. To this end, Random Forest (RF) and Gradient Boosting Machine (GBM) classifiers were used for forecasting customer behavior using RFM analysis on an online retail dataset for identifying the optimal month for forecasting customer behavior by training the classifiers on two months of data and testing on one month.

3 Methodology

3.1 KDD Process for Customer Segmentation

Knowledge Discovery from Databases (KDD) is an approach to systematically extract valuable insights from large and complex datasets. There are several steps in this process, each with its own goals and role to transform raw data into actionable information. This section describes the KDD process illustrated in Figure 1, which performs customer segmentation on an online retail dataset.

Target data selection: The first step of the KDD process is to choose what data to work with. For customer segmentation, this means selecting an online retail dataset that contains customer ID, purchase history and product details among others. By doing so, only those features will be considered which can provide useful insights into customer behavior.

Data preparation Data cleaning is an essential part of any project where data preprocessing plays a huge part in the model training and evaluation. This involves dealing with missing values, removing duplicates if necessary, combining columns if appropriate and other steps that help ensure that no transaction relating to a given person or time period is left out.

Calculation of RFM Score: RFM (Recency, Frequency, Monetary) score calculation serves as a way of measuring customer behavior. It assigns scores based on three factors for every client: Recency - the amount of time since the customer's most recent purchase; Frequency - the number of purchases made by the customer during a certain period, for instance, a month duration in case of this research; Monetary - total money spent by each buyer over a specified duration and is critical in identifying the high profile customers for the company's business. These scores are crucial for the segmentation process and assists in understanding the purchasing behavior of customers.

Cluster Creation: Clustering techniques like K-means clustering can be used for dividing the whole customer base into separate segments with similar patterns based on RFM-scores. This plays a huge part in the decision-making process of businesses to figure out the target customers using personalized marketing promotions.

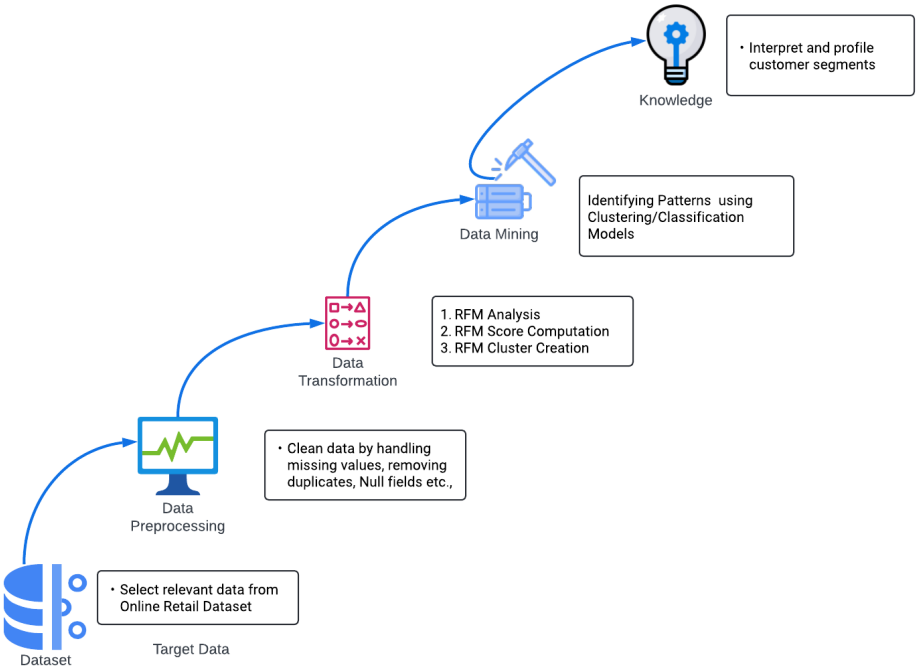

Figure 1: KDD Process Methodology for Customer Segmentation

Classification Models: RF(Random Forest) and GBM(Gradient Boosting Machines) are classification algorithms that can be employed to forecast clusters for new customers based on their attributes. These models assign different types of records to existing customer groups using RFM scores.

Evaluation: The evaluation phase examines how well the developed classification models and resulting clusters meet the research objectives based on the training data. Moreover, it should also check if segments derived from them are meaningful and align with organizational strategies. If they don't, then there might be need for different segmentation approaches that necessitates to go through the previous KDD phases.

Knowledge: The final stage comprises interpretation and profiling of customer segments where by analyzing each segment's attributes so as to extract knowledge and give recommendations. Businesses together with their marketing teams can therefore utilize this information in developing consumer engagement plans by means of discounts, offers, and promotional campaigns.

In conclusion, KDD often requires going through several iterations across the phases so that extracted information becomes accurate and useful for decision making.In addition to that, businesses can gain deeper insights about their clients' needs through applying KDD approach on customer segmentation for online retail dataset and enable them to engage effectively via personalized marketing.

4 Design Specification

Customer analytics studies focus on handling million of records. A company however might wonder whether the extensive data required for their studies is necessary. Think, for instance in buying 3rd party data, how much data should they obtain to perform

customer segmentation? How many months of data should we use to be confident on our LTV predictions? This research addresses those question. We present a detailed study of the best month we can predict. Moreover, what algorithm should we use for customer segmentation? In this research we used RF and GBM, and also evaluated the need of hyper-parameter optimisation and autoML.

5 Implementation

5.1 Dataset Analysis

The various elements that make up the design characteristics of the research are, for instance; dataset, RFM analysis and classification. Each of these components is important to the realization of the objectives aimed at finding out which month is best for consumer behavior forecasting as well as solving the problem of lacking "good" customers in some months. The Online Retail Dataset used in this study contains transactional data from an UK based online retailer. It starts on December 1st, 2010 and ends on December 9th, 2011 with 541,909 rows having following attributes:

- InvoiceNo: Unique identifier per transaction
- StockCode: Unique identifier per product
- Description: Product Name
- Quantity: Number of items bought in a transaction
- InvoiceDate: Date and Time when a specific item was bought in a transaction
- UnitPrice: Per Unit Product Price
- CustomerID: Unique identifier per customer
- Country: Customer country of residence

Before beginning RFM analysis, it is necessary to preprocess the dataset so as to ensure that data quality and uniformity is achieved. This involves dealing with null values such as rows without customer ids , price values without item quantity etc., dropping duplicate records, and converting data types wherever necessary. For example, Customer ID and Quantity as integers and InvoiceDate into datatime object. A new column TotalPrice is created by multiplying Quantity and UnitPrice columns.

5.2 RFM Analysis

RFM analysis is performed to figure out consumer behavior and categorize them based on their buying habits. This study works with Online Retail Dataset for applying RFM analysis and calculating RFM score of each customer. The 'R' value is calculated by getting the difference between maximum invoice date of the dataset and most recent invoice date for every customer. The smaller the value, the better because it indicates that a purchase has been made recently. 'F' value is derived from number of purchases made by a customer during specific period.The higher 'F', better its score will be. The 'M' value is calculated by taking sum of 'TotalPrice' spent over some time frame by all

customers. A higher value of 'M' indicates that the customer is a prolific spender. The RFM score represents an index reflecting Recency-Frequency-Monetary Value rankings or weights given according to different components such as R (recency), F (frequency), and M(monetary). A formula can be devised to calculate how likely customers are going to repurchase depending on their last time buys as well as how often they make purchases in general and how much they spend altogether within a particular period. It can be represented mathematically as

$$0.15Rscore + 0.28Fscore + 0.57Mscore$$

The weights were chosen such that Recency gets weight 15%, Frequency receives 28% and Monetary Value takes up most with 57%. The values for R, F, and M scores are determined through weighting or ranking mechanisms but these can vary depending on the business context. However certain standard recommendations have been proposed like assigning lower values if higher numbers are needed indicating better quality services offered by the business. The weights correlate inversely so that a decrease in one component will lead to an increase in another. Thus small weighted RFM (RFM score: '111') implies big spending which means more frequent visits made recently (lower R_score), higher purchase frequency (higher F_score) accompanied with greater amounts spent on goods or services (higher M_score).

5.3 Classification Models

This section discusses in brief the ML classifiers,RF and GBM and the rationale behind their choosing.

Random Forest (RF): An ensemble learning technique that trains multiple decision trees and returns the most frequent class for classification or average prediction for regression of individual trees. RF was chosen for its salient characteristics including the ability to handle high dimensional data, avoid overfitting, and feature importance rankings. RF was employed in this study to segment customers into different categories based on their RFM scores. The RF classifier will undergo training on a two-month window period of the online retail dataset and then be evaluated using a one-month dataset to determine the optimal month for predicting consumer behavior. The RF classifier's performance will be assessed using accuracy, precision, recall, and F1-score.

Gradient Boosting Machine (GBM): GBM is a type of learning method that combines weak learners—usually decision trees— to form one strong learner. What happens in GBM is every new model which is included in the ensemble attempts correcting errors committed by previous models sequentially. It is widely known for having excellent predictive accuracy and being able to deal with complex feature interactions effectively as well. The GBM classifier to segment customers into different groups based on their RFM scores. The GBM classifier will also be trained using two months' data and tested against one month's data to identify which month gives the best prediction about consumer behavior. The performance of the GBM classifier will be compared with that of the RF classifier through this study to establish which model performs better.

The rationale behind the selection of RF and GBM is derived from the fact that both are ensemble models with the ability to handle higher dimensional data and prevent ovefitting on complex features. The workflow depicted in Figure 2 was coded using Python, supplemented with other data manipulation and analysis libraries (Pandas, Numpy, Matplotlib). In addition to this, scikit-learn was used to implement both RF as

well as GBM models. The code implementation and deployment was done using CoLab – an interactive coding environment powered by Jupyter Notebooks – hosted on Google's Cloud datascience platform.

After installing and import the necessary libraries required for the execution, the dataset can be loaded from the archive [1] as a pandas dataframe.

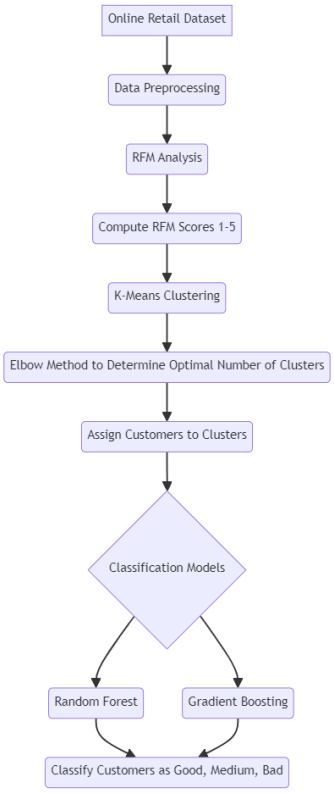

Figure 2: Customer Segmentation – Design Workflow

5.4 Data Preprocessing

The retail dataset is processed further with only the necessary features to compute the RFM score and perform RFM analysis later to make it more appropriate. Some of the preprocessing done on the dataset are: dropping all those 'CustomerID' fields with NULL rows, changing the data type of 'CustomerID' and 'Quantity' to integer type using 'astype(int)' function, convert 'InvoiceDate' column to a DateTime format (mm/dd/yyyy, hh:mm:ss) using the pandas function, pd.to_datetime(). A duplicate column of 'InvoiceDate' is generated and named as 'InvoiceDate2' to help with the selection of different date windows during the model training. Finally, a new column, 'TotalPrice' is created by multiplying the 'Quantity' by 'UnitPrice' and grouped by 'CustomerID'.

5.5 RFM Score Calculation

The percentile ranks are computed to generate the RFM score of each customer and divided into equal-sized bins that represent the levels using pandas routine, pd.qcut. A

[1] `https://archive.ics.uci.edu/ml/machine-learning-databases/00352/Online%20Retail.xlsx`

score is assigned to each bin in levels of 1 to 5 for each R, F and M. The combined RFM score is generated by simply stringing together the individual values of R, F, and M. For example, if R score is 2, F score is 4, and M score is 1, then RFM score will be "241". The weighted RFM score is obtained by

$$rfm\_score = 0.15 * r\_score + 0.28 * f\_score + 0.57 * m\_score \quad (1)$$

It works by allocating the pre-established weights to each component and combining them to generate a composite number for each customer. The usage of other weights is also possible and it's just a choice of which of R, F, and M matters most to the company.

In this research, a lower weighted RFM score points to a 'Good' costumer, since it implies there were more recent purchases (lower R_score), high purchase frequency (lower F_score), and more spends per purchase (lower M_score). Thus, '111' would represent the most ideal 'good' customer, while '555' would point to a 'bad' customer.

5.6 Segmentation using percentiles

We took the approach to segment the weighted rfm score using a percentile approach as in Table 1

Table 1: Summary statistics of the weighted RFM score, Eq.1

count	4322
mean	2.993552
std	1.252473
min	1
0%	1
25%	1.99
50%	3
75%	4.13
max	5

The average RFM score across all customers is 2.993552, as stated in Table 1. The percentile value at 0% is defined as the lowest value of 1, similarly for 1.99 for 25%, 3 for 50%, 4.13 for 75%, and 5 for 100%. Regarding this, a little dispersion has been shown by RFM scores towards lower numbers. This can be proven by median (3) positioning near minimum (1.0) but not maximum (5.0) values as well as standard deviation equaling to 1.25 which tells about moderate scattering from average – hence acceptable diversity among customer behavior patterns.

From these observations, we could take the following thresholds to categorize customers into three classes ('g', 'm', and 'b') based on their weighted RFM score with the following logic: A score between 1-2 belongs to 'g' good customers, 2-4 belongs to 'm' medium customers, and 4-5 represents a 'b' bad customers.

The next step in the implementation process is to find the answers to the research questions presented in Section 1. This is achieved through model training using RF and GBM using RFM scores as the input and Class as the output parameter to be predicted. The next section discusses this in detail considering the three research questions as case studies.

6 Evaluation

The results are evaluated by following a sliding window approach for selecting the months for training and testing, by considering 2 months for training and 1 month for testing at any point of time. This will help us to analyze the dataset more closely and try to find out the best month for forecasting. The dataset starts from 01 December 2010 and ends on 09 December, 2011. By following this sliding window approach, we should be able to make forecasts for a 10-month period.

6.1 Case Study 1: Best Month to Forecast using RF

We initialise the RF classifier with a maximum depth of 2, meaning that the depth of each tree is 2 levels, which can help prevent overfitting and reduce the complexity of the model. From training the model and analysing the classification report for 10-month forecasts, it can be deduced that only a few 'good' customer classes occur across the selected period as seen in Table 2. This indicates that the classification is not done correctly, or the samples are too less or imbalanced to make proper class predictions. The total number of samples being less for the months starting from Feb – August is a learned fact from Table 2 which may be the reason for the inaccurate predictions of customer classes during that period. The customer classes are more balanced during the period from Sep to November, resulting in better predictions. This tells us that there is a possibility that customer classification can be improved by changing the RFM score calculation function, Eq.1, or by widening the threshold value for good customers. This needs to be evaluated before arriving at any concrete conclusions. But with the present approach and results, it can be presented that the best month for forecast may be 'April' evident from the weighted average F1-score and the worst month for forecast is November, with a score of 0.813, which is less.

Table 2: Support Table for Customer Class

Month	bad	good	medium
Feb	67	2	31
Mar	118	2	44
Apr	100	1	36
May	89	5	66
Jun	90	5	86
Jul	83	9	83
Aug	79	14	123
Sep	119	33	242
Oct	149	113	379
Nov	128	437	704

The weighted average F1 score calculates the F1 score for each class independently, but when it averages the scores, it weighs them by the number of support instances in each class. This provides a more realistic measure of the model's performance compared to the macro average F1 score, which gives equal weight to all classes regardless of their support, providing a more accurate reflection of the model's overall performance by considering the

class distribution. The weighted average F1 scores, are plotted for the period, February to November as shown in Figure 3.

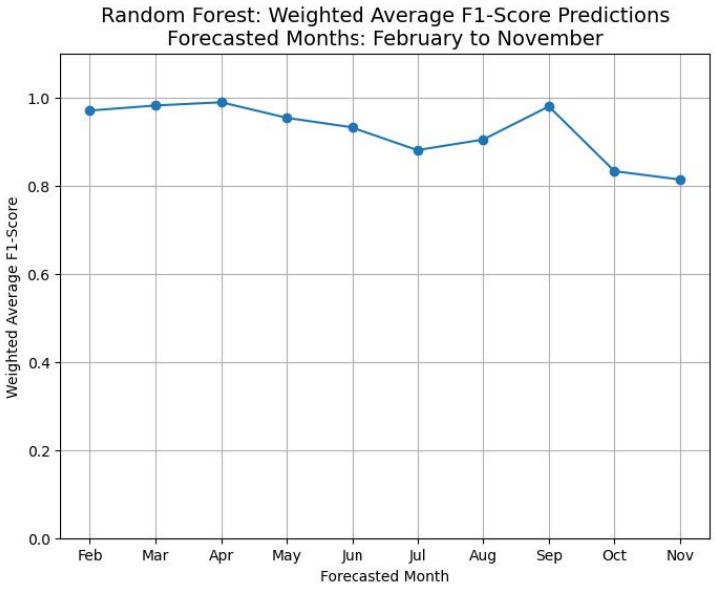

Figure 3: Weighted Average F1-Score for RF. Here, April is the 'best' month and November is the 'worst' month for forecasting customer behavior.

6.2 Case Study 2: Best Month to Forecast using GBM

GBM classifier is defined as GradientBoostingClassifier(n_estimators=100, max_depth=1, random_state=0) with n_estimators parameter set to 100, which specifies the number of decision trees in the ensemble. Increasing the number of estimators can improve performance but also increases training time. The max_depth parameter is set to 1, indicating that the decision trees will be shallow (stumps). Shallow trees help prevent overfitting and can be beneficial when combined in an ensemble. The same analysis presented for RF regarding the poor 'good' customer class for most of the months during the test period is also valid here. From the predictions, it can be seen that the best month for forecast may be 'February', evident from the weighted average F1 score and the worst month for forecast is 'November', with a score of 0.9269. The weighted average F1 scores, are plotted for the period, February to November as shown in Figure 4.

6.3 Case Study 3: Stratified Classification using RF and GBM

Upon analysis, we realised that we have low number of "good" customers in some months. It does not do any good to the classification model. Those months with the low number of "good" customers need to be considered as 'invalid' as it points to some abnormality either in the way of the number of samples being less or can represent an imbalanced dataset. Since the months have so much disparity in the number of "good" customers, the use of a stratified classifier for forecasting using a 10-CV validation can present a better solution for improving the classification performance. Utilizing a stratified data sampling, like *StratifiedKFold* in Sklearn, can enhance the precision of a model by guaranteeing that

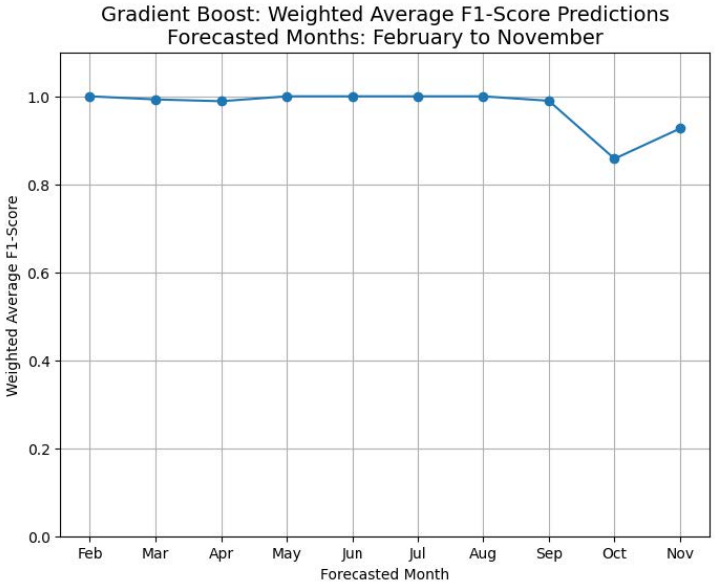

Figure 4: Weighted Average F1-Score for GBM. Here, February is the 'best' month and November is the 'worst' month for forecasting customer behavior.

each fold in the cross-validation procedure contains a proportional distribution of the target classes. This is especially beneficial when working with imbalanced datasets, in which certain classes have a notably lower number of samples compared to others.

Stratified classification is used in imbalanced datasets to ensure that each fold contains a sufficient representation of the minority classes, which have much fewer samples than the majority classes. This mitigates the potential bias of the model towards the majority class and enables it to effectively capture the patterns exhibited by the minority classes. In our analysis, by using StratifiedKFold with n_splits=10, 10 stratified folds are created for cross-validation. Each fold will have a similar distribution of the target classes as the original dataset. The cross_val_score function then evaluates the classifier's performance using these stratified folds, providing a more accurate estimate of the model's performance. The performance of the 10-fold stratified classifier is measured in terms of accuracy and the median accuracy for the 10-folds are calculated as shown in Figure 5 for both RF and GBM. With the use of the stratified classifier for RF and GBM, GBM model outperformed RF with a classification accuracy of 98.497% compared to 90.74% of RF. This also negates the need for performing hyper-parameter optimization using AutoML.

6.4 Discussions

Use case 1

From use case 1 (aligned to Research Objective 1), we noticed the number of "good" customers is not even. We have very little numbers during February to September. However, good numbers of "good" customers only appear in October and November. We observed in individual runs of the classifier, that high accuracy was found, but the classifier missed the "good" customers. In other words, the classifier excelled to separate medium and bad customers, but weak in finding good ones.

If a company wants to use this classification, the only models that are of practical

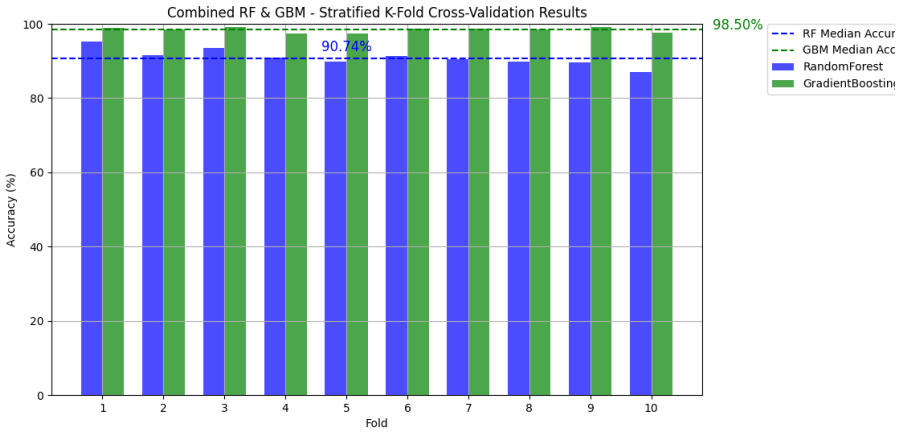

Figure 5: Comparison of Stratified 10-fold Cross-Validation Accuracy - RF vs GBM. GBM is constantly better than RF.

interest would be the October and November. Another possibility, is to change the thresholds we set for good, medium, and bad customers. Let's say, for instance, the "good" customers could be taken from 1-3, instead of 1-2.

From a company perspective, the weights assigned to Eq. 1 could be changed to other weights. It might reflect a focus on customer frequency, rather than purely monetary value.

In those months, as in Figure 6 we can also see the number of active customers increases towards the end of year. It is good insight for staffing and stocking inventory, and computing resources. For example, tuning up the load-balancers and spin new instances in different geographies. We could also have created only two models, one to forecast from Feb-Aug, and another one from Sep-Nov. The raw data can be seen in Table 2.

From Figure 3, we can see all predictions are very high. This can be due to the size of the dataset, and problem complexity. Or, perhaps RF is very good indeed for this classification task. The highest accuracy is found in April, while the lowest is November. RF is a tree-based classifier, we set the max depth to 2, but this is a value we can adjust according to the dataset. Perhaps a single decision tree might be just sufficient.

Use case 2

For use case 2 (aligned to Research Objective 2), we can see from Figure 4 that the GBM trend is in line with RF results. It has an upward trend toward the end of year, while remains very high during the rest of year. It shows steady predictions from February to September, compared to RF. The classification accuracy remained very high regardless of the sample size in each month.

We noticed, as in the case of RF, some of the predictions are very high using F1 score, even if one of the classes ("good" customers) gets zero classification accuracy. The overall F1 score is exceptionally high. We propose GBM to be the algorithm of choice when the datasets become larger or more complex.

A note of caution: we based our findings and discussions from one single customer dataset. A larger dataset could however be analysed using the proposed methodology. The data analysis pipeline should be applicable out of the box.

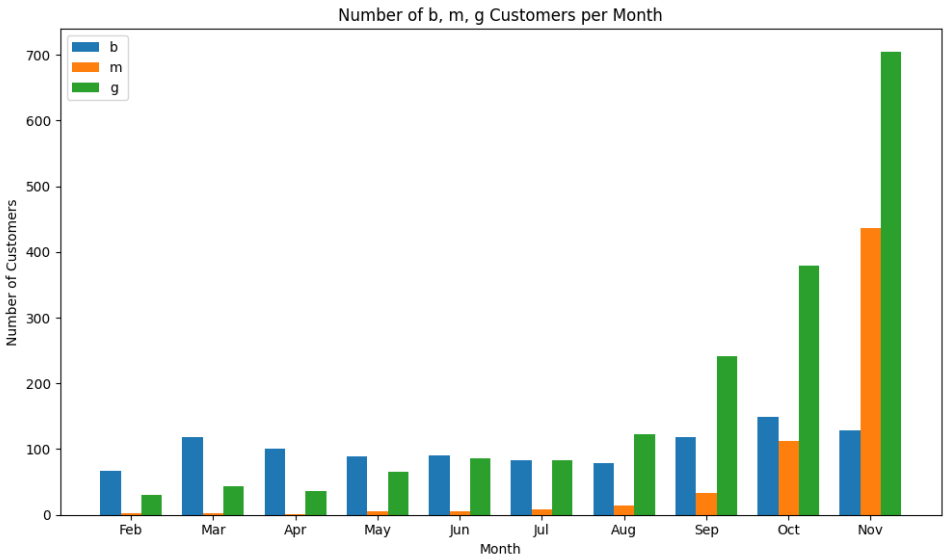

Figure 6: Active Customers Trend. The increase in the number of active customers towards the end of the year for b:bad, g:good, and m:medium classes is evident from the plot. Our forecast models are particularly important from September, as the largest share of customers appear from that point.

Use case 3

Finally, for Research Objective 3, for the last use case, we used a stratification. Good, medium, and bad customer classes are equally sampled before classification. Both classifiers performed well across the 10-fold cross-validation steps. We used the standard 10-folds to calculate an average F1 score, median, and confidence intervals, and the performance of Stratified K-Fold CV are shown in Figure 5 for both RF and GBM. From Figure 5, we can infer that GBM outperforms with a median score of 98.50% compared to RF with a score of 90.74%.

As for Research Objective 4. We noticed the F1 accuracy of this stratified classification to be very high. So, we think hyper-parameter optimisation and autoML is not required at the moment, but could be of interest for future work, especially for RF.

7 Conclusion and Future Work

This research investigated the use of RFM analysis and machine learning classifiers, for customer behavior segmentation using an online retail dataset. In comparison to existing literature, this research focuses on identifying the best month for forecasting with limited data as this is crucial for SMEs with smaller datasets or smaller budgets for the purchasing of 3rd party data. We all know seasonal variations in our shopping behaviour. Thinking of those, we tried to identify the best month for short-term forecasting. We segment our customers and trained a classifier on two months of data and one month for testing. We noticed this strategy could lead a class imbalanced problem. The issue of having few "good" customers in certain months. Broadly speaking, the segmentation can help SMEs to target those "good" customers throughout the year. In the beginning of the year can entice them to remain loyal, while you know they will come with a spending

spree towards the end of the year. The research methodology presented allows SMEs to find those "good" customers and their cut-off date.

The key findings of the study are as follows:

- Using the RF classifier, April was identified as the best month for forecasting customer behavior based on the weighted average F1-score, while November was the worst month.
- The RF was outperformed by the GBM classifier, with February as the best month for forecasting and November as the worst month as can be seen from the weighted average F1-score.
- The classification performance was negatively affected by a few "good" customers in some months which indicates data abnormalities or imbalance. Handling the class imbalance of few good customers in certain months due to lower number of customer sample data available for those months. This issue was addressed using the stratified cross-validation approach.
- As regards to classification accuracy, the RF and GBM classifiers with 10-fold cross-validation proved to be more effective in their stratified structure than before. This resulted in an improvement of about 98.49% for GBM and 90.74% for RF, thus making hyper-parameter optimization unnecessary.

Future Research

Our research can be extended in a number of ways, for instance:

- Modify the RFM score calculation function or change the threshold value for "good" customers and see its effects on classification performance and forecast accuracy.
- Different customer segments should be considered. In this study, a commonly employed RFM analysis was used whereby R, F, and M columns are divided into five equal parts based on percentiles but other methods exist too as well as different class divisions can be attempted.
- Preprocessing techniques like SMOTE that over-samples the minority class, or adjusting class weights could have been utilized to help models learn patterns from all classes properly.
- ML classifiers such as Support Vector Machines (SVM), Neural Networks can be used in conjunction with RF and GBM for customer behavior prediction performance assessment.
- More customer characteristics like demographics or product categories can be included to the analysis for good understanding of customer behaviour.

References

Anitha, P. and Patil, M. M. (2022). Rfm model for customer purchase behavior using k-means algorithm, *Journal of King Saud University-Computer and Information Sciences* **34**(5): 1785–1792.

Ansari, A. and Ghalamkari, S. (2014). Segmenting online customers based on their lifetime value and rfm model by data mining techniques, *International Journal of Information Science and Management (IJISM)* pp. 69–82.

Chen, Y.-L., Kuo, M.-H., Wu, S.-Y. and Tang, K. (2009). Discovering recency, frequency, and monetary (rfm) sequential patterns from customers' purchasing data, *Electronic Commerce Research and Applications* **8**(5): 241–251.

Chiliya, N., Herbst, G. and Roberts-Lombard, M. (2009). The impact of marketing strategies on profitability of small grocery shops in south african townships, *African journal of business management* **3**(3): 70.

Christy, A. J., Umamakeswari, A., Priyatharsini, L. and Neyaa, A. (2021). Rfm ranking–an effective approach to customer segmentation, *Journal of King Saud University-Computer and Information Sciences* **33**(10): 1251–1257.

Dam, N. A. K., Le Dinh, T. and Menvielle, W. (2021). Towards a conceptual framework for customer intelligence in the era of big data, *International Journal of Intelligent Information Technologies (IJIIT)* **17**(4): 64–80.

Dogan, O., Ayçin, E. and Bulut, Z. (2018). Customer segmentation by using rfm model and clustering methods: a case study in retail industry, *International Journal of Contemporary Economics and Administrative Sciences* **8**.

Dursun, A. and Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of rfm analysis, *Tourism management perspectives* **18**: 153–160.

Ernawati, E., Baharin, S. and Kasmin, F. (2021). A review of data mining methods in rfm-based customer segmentation, *Journal of Physics: Conference Series*, Vol. 1869, IOP Publishing, p. 012085.

Frasquet, M., Ieva, M. and Ziliani, C. (2021). Online channel adoption in supermarket retailing, *Journal of Retailing and Consumer Services* **59**: 102374.

Gustriansyah, R., Sensuse, D. I. and Ramadhan, A. (2017). A sales prediction model adopted the recency-frequency-monetary concept, *Indones. J. Electr. Eng. Comput. Sci* **6**(3): 711–720.

Gustriansyah, R., Suhandi, N. and Antony, F. (2020). Clustering optimization in rfm analysis based on k-means, *Indonesian Journal of Electrical Engineering and Computer Science* **18**(1): 470–477.

Handojo, A., Pujawan, N., Santosa, B. and Singgih, M. L. (2023). A multi layer recency frequency monetary method for customer priority segmentation in online transaction, *Cogent Engineering* **10**(1): 2162679.

Hosseini, M., Abdolvand, N. and Harandi, S. R. (2022). Two-dimensional analysis of customer behavior in traditional and electronic banking, *Digital Business* **2**(2): 100030.

Huang, Y., Zhang, M. and He, Y. (2020). Research on improved rfm customer segmentation model based on k-means algorithm, *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)*, IEEE, pp. 24–27.

Jasek, P., Vrana, L., Sperkova, L., Smutny, Z. and Kobulsky, M. (2018). Modeling and application of customer lifetime value in online retail, *Informatics*, Vol. 5, MDPI, p. 2.

Kumar, S. J. and Philip, A. O. (2022). Achieving market segmentation from b2b insurance client data using rfm & k-means algorithm, *2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, Vol. 1, IEEE, pp. 463–469.

Monalisa, S., Nadya, P. and Novita, R. (2019). Analysis for customer lifetime value categorization with rfm model, *Procedia Computer Science* **161**: 834–840.

Rungruang, C., Riyapan, P., Intarasit, A., Chuarkham, K. and Muangprathub, J. (2024). Rfm model customer segmentation based on hierarchical approach using fca, *Expert Systems with Applications* **237**: 121449.

Sebald, A. K. and Jacob, F. (2020). What help do you need for your fashion shopping? a typology of curated fashion shoppers based on shopping motivations, *European Management Journal* **38**(2): 319–334.

Shokouhyar, S., Shokoohyar, S. and Safari, S. (2020). Research on the influence of after-sales service quality factors on customer satisfaction, *Journal of Retailing and Consumer Services* **56**: 102139.

Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S. and Rahmani, R. (2018). Customer segmentation and strategy development based on user behavior analysis, rfm model and data mining techniques: a case study, *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*, IEEE, pp. 119–126.