An Eclectic Approach for Predicting Customer Segmentation to Facilitate Market Basket Analysis

MSc Research Project
Data Analytics

Ashwini Mohan
Student ID: x19220618

School of Computing
National College of Ireland

Supervisor: Prof. Athanasios Staikopoulos

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ashwini Mohan
Student ID:	x19220618
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Prof. Athanasios Staikopoulos
Submission Due Date:	31/01/2022
Project Title:	An Eclectic Approach for Predicting Customer Segmentation to Facilitate Market Basket Analysis
Word Count:	7659
Page Count:	24

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:	Ashwini Mohan
Date:	31st January 2022

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):	

An Eclectic Approach for Predicting Customer Segmentation to Facilitate Market Basket Analysis

Ashwini Mohan
x19220618

Abstract

Data mining technologies are being applied to large amounts of consumer data to gain useful insights about customers, implement customer segmentation for marketing purposes, understand individual purchasing behaviour, create recommendation systems, and so on, all of which, when implemented, can help the organization maximize its profitability. The primary objective of this research was to **combine machine learning with data mining techniques, to forecast customer segments based on the transaction data**. The secondary objective of this **study is an extension of the earlier research and a novel approach** conducted in the same domain and seeks to **evaluate the hypothesis that the application of association mining rule on segmented customer data is more effective** compared to its application on the entire dataset. An Online retail dataset II having 1,067,371 transactional records was used in this research, which was divided into 4 clusters based on the RFM analysis and K-Means clustering technique. The initial objective of this research was tested against four classification models, KNN, Decision Tree, Light Gradient Boosted Machine (LGBM) to perform multiclass classification. On the basis of training time and K-Fold Validation accuracy, **Random Forest was deemed as the best model** with a K-Fold test accuracy of 85.15% and training time of 0.1935 seconds. The results obtained for the second objective of this research led to **rejecting the hypothesis**, since no recommendations were generated for two of the classes. The approach undertaken in this research is effective in categorizing customer to their respective groups, producing a satisfactory outcome and provided information about the best practices to embed in the topic to get better results.

Keywords: Association Mining Rule, Decision Tree, K-Means Clustering, KNN Classifier, LGBM, Random Forest, RFM Analysis

1 Introduction

Big data processing helps enterprises analyse huge amounts of data in order to acquire business insights, comprehend trends, important patterns about their customers' purchasing habits, rival product evaluations, and so on, resulting in decreased risk. As a result, classification and association mining have become two key data mining approaches that are extensively applied across all sectors. Separate research in the retail business has been conducted that utilizes data mining approaches to categorize clients in order to anticipate potential transactions, as well as the utilization of association mining algorithms to discover market trends(Azuaje; 2006).

Additional challenges that a consumer may have during their online buying experience include getting promotions or offers that are irrelevant to them. This is a downside for both the merchant and the customers, as, from an organizational viewpoint, the offers are not reaching the proper customers and sales are not growing as anticipated, and consumers may be upset as a result of not obtaining any relevant offers or discounts. The fundamental source of this issue can be that the consumer segmentation technique applied is imprecise, or that the algorithm used to segment consumers does not have high accuracy, leading marketers failed to fulfill customers' expectations. Furthermore, from the viewpoint of a retailer, identifying the interdependence of many items can be advantageous as they could then promote relevant items to consumers, resulting in improved sales or by delivering a combined promotional offer to improve a product's sales. This can be accomplished through market basket analysis or association mining rules, in which interesting product interrelationships are examined through customer purchase history. Association algorithms, such as the Apriori Algorithm, and the FP-Tree Algorithm (Frequent Pattern–Tree Algorithm), may be used to detect the link between item-sets. (Kaur and Kang; 2016). In the literature review part, the algorithms are studied in detail, as is why the Apriori Algorithm is relevant for this study.

Several studies(Tsai and Chiu; 2004) (Chan; 2008) (Camilleri; 2017) have been conducted to demonstrate the benefits of maintaining customer relationship management and how marketing techniques tailored to each consumer segment improve customer experience and ensure customer loyalty. There have also been some research (Aguinis et al.; 2013)(Kulkarni; 2012) that used association mining techniques to determine the interrelationship of products or item sets across domains. Overall, implementing both the segmentation and association mining rules has seen a positive result. The state-of-the-art is that the RFM and K-Means Clustering algorithms have been proven to segment the customer more accurately compared to other data-mining technologies such as Fuzzy C-Means, Repetitive Median K-Means. However, the application of association mining rule on customer segment to recommend consumer relevant product is yet to be conducted (Christy et al.; 2018). In this context, the main goal is to early categorize customers into segments based only on the consumers' purchasing behaviour by applying RFM analysis and the K-Means Clustering method and developing a multiclass classification model to then forecast the consumers. It may be claimed on higher prediction efficiency and on the grounds that RFM and K-Means clustering accomplishes the segment, thus requirement of multiclass classification is not necessary. But application of K-Means is still computationally expensive and has its drawbacks. Also, the applicability of association mining rule or market basket analysis on segmented clients is yet to be explored. Specifically, this research addresses the following questions:

1. How feasible is machine learning methodology when integrated with the RFM and K-Means Clustering to facilitate in predicting multiple customer segments?
2. How effective is the association mining rule when applied to customer segments for targeted product recommendations?

The **primary objective** of this study is to establish a high performing multiclass classification model, and the **secondary objective** is to test the hypothesis that association mining rule application on segmented customer data is more effective compared to its application on the complete dataset.

This research will aim to integrate both strategies into a retail transaction dataset by first segmenting the data using the RFM methodology and then attempting to develop an ideal multi-class classification model that will guarantee that the customers are appropriately categorized into their respective classes. Lastly, market basket analysis will be conducted on the segmented consumers to undertake a more targeted study that might yield better results.One of the limitations of this research is that the proposed approach is tested on transactional data, and hence the approaches applied here may not be relevant to datasets comprising clients demographical or physiological details.

This paper is arranged as follows : Section 2 offers a full review of the literature, with an emphasis on the research undertaken on the stated subject of discussion in Section 1. Section 3 discusses the procedure followed and data mining technologies that will be utilized to address the research question. Section 4 elaborates on the architectural framework that underlies the implementation. Section 5 discusses the models constructed, tuning required for each model and data transformation undertaken. Section 6 addresses the outcomes and limitation of this research approach. Section 7 concludes the findings and describes future scope.

2 Related Work

As business data systems make massive volumes of data accessible, data mining technologies emerge as logical solutions to the aforementioned challenges. Several reviews have been used to evaluate performance in the retail and customer relationship management sectors (Ngai et al.; 2009; Saxena et al.; 2017; Raju et al.; 2007). These reviews feature multiple examples of the usage of different data mining approaches to help with a broad variety of retail-related issues. As a result, solutions such as knowledge discovery (KD) and data mining (DM) have been offered (Hiziroglu; 2013). This section delves into the cutting-edge of data mining and machine learning techniques for segmenting customers, predictive segmentation, and association techniques for trend analysis.

This research review is divided into three sections: customer segmentation approach, multi-class categorization model for predictive segmentation, and market basket analysis. The purpose of this literature review will be to identify clustering algorithms to be applied in this research. Furthermore, different multiclass classification models will be found for use in this research in order to undertake a comparative analysis and choose a model that is optimal from all perspectives and suited for future prediction. A thorough investigation is also performed to decide which approach would be most suited for market basket analysis.

2.1 Customer Segmentation Approach

This section emphasizes the increasing importance of knowing customer behaviour, and how it could benefit the business, which is the motive behind this proposed study.

The first studies on association and classification demonstrated the use of data mining technique in CRM personalized marketing and loyalty techniques, signalling a rise in consumer segmenting research. (Ngai et al.; 2009; Saxena et al.; 2017) began investigating how various categorizing models are used for multiple classification strategies such as hierarchical, partitioning, model-based classification, and so on. The user purchase transaction is classified as unsupervised learning since it groups objects together based on an unknown pattern.Despite the availability of more statistically advanced approaches, according to (Verhoef; 2003), RFM remains the second best prevalent method employed by advertisers. The RFM method evolved from catalogue marketers' informal recognition that three criteria, Recency, Frequency, and Monetary, are commonly the best prospects for new product offerings.

McCarty and Hastak (2007) compared RFM, CHAID, and logistic regression on two datasets (multi-division mail order and member contribution to their non-profit organization) by analyzing the results at multiple depths of files. They applied RFM model by quantifying customer actions using their behaviour quantile approach. A quantile is a collection of records that reflects 20% of the data. It has the benefit of having an equal number of consumers in each group. This idea is followed by a customer quantile in segmenting consumers evenly by differentiating them with defined RFM metrics and scores (Hosseini et al.; 2010). However, if a significant proportion of consumers only bought once in the frequency measurement, segmentation using the quantile technique may be skewed. As a consequence, two or three quantiles of consumers may exhibit similar behaviour. To solve the sensitivity issue of the customer quantile technique, the behaviour quantile method organize customers by creating arbitrary cut-offs on behavioural percentage (Dursun and Caber; 2016). It provides the benefit of grouping clients who exhibit similar behaviour. The experiment was evaluated based on the reliability of the segmentation approach, where all three models provided accurate prediction and gain percentage of the segmentation method, in which RFM and CHAID models performed significantly better than logistics regression; however, the RFM model had an advantage because its performance was 10% higher than CHAID when the response rates were high. This article does bring up the point that RFM concentrates on transactional information while ignoring individual characteristics such as demography and psychology, which are equally crucial for determining individual preferences.

Hu and Yeh (2014) later studied transactional data through the RFM pattern and a novel method for discovering whole sets of respective patterns to reproduce client sets by combining RFM analysis with frequent pattern mining. Rather than evaluating pattern values from the customer's perspective, their study compared rankings by using frequent patterns RFM metric. Cheng and Chen (2009) used DM methodologies to create a new customer segmentation strategy based on RFM, demographic, and LTV data (Namvar et al.; 2011). Mishra et al. (2017) carried out a study to compare the segmentation approaches, including K-Means, FCM, subtractive, and mountain. The algorithms showed efficiency in varying conditions, and K-Means outperformed other algorithm in terms of accuracy, but showcased poor result when implemented in high-dimensional data space. Another limitation of this approach is that it does not handle outliers efficiently and is susceptible to unrelated data. To address the inadequacies of previous research, An RFM attribute combined with K-Means clustering was used in a later study in the electronics industry to impartially segment customers (Cheng and Chen; 2009). The investigations demonstrated that the RFM and feature-based K-Means segmentation approach outperforms all the other models. Also, Christy et al. (2021) did a comparison study between RFM analysis with K-Means, Fuzzy C Means, and Repetitive Median K-Mean, concluding that RFM with K-Means was ideal to implement for transactional segmentation because it consumed less time and required fewer repetitions than the other techniques, making it an appropriate choice for this research.

According to the reviewed literatures, consumer segmentation based on RFM analysis combined with K-Means clustering technique outperforms other techniques and was shown to be the best fit for dealing with transactional data. The RFM approach has evolved over time by broadening its scope to include client demographics and psychological elements. Since the dataset used in this research is behavioural in nature and lacks customer information or attributes, RFM analysis with K-means clustering will be used to perform customer segmentation in this study.

2.2 Multi-class Classification Model For Predictive Segmentation

This section focuses on the data mining research undertaken to accomplish predictive clustering using multi-class classification models.

According to Aly (2005), the goal of a multi-class classification algorithm is to distinguish each sample using a unique label, which will then be trained using other attributes linked to each label to create a learning pattern that will aid in understanding the new data and classifying them in their respective samples. The goal of supervised multiclass classification algorithms is to give a class label to each input sample (ibid, p.1). The author also conducted a survey on multiclass classification models by conducting research on binary classification, identifying the algorithms primarily used for binary classification, and then explaining how certain binary class classifiers such as Decision Tree, K-Nearest Neighbour, Support Vector Machines, and others can be extended to perform multi-class classification. Using binary decision trees, a tree model with a ”True–False” structure may be developed to split data into multiple groups depending on their respective features (Olson and Chae; 2012).Simulated neural networks (SNN) are particularly well-suited for non-linear relationships in general, but their efficiency decreased as they cannot adapt to new data. Logistic regression models, on the contrary, are easy to adjust to any new information, however, class cut-off might be difficult (Olson and Delen; 2008). Bagging, boosting, and randomizing were employed on a decision tree model in one of the experiments to examine the performance of the varied techniques on a model, and were assessed based on noise variance ((Dietterich; 2000). In scenarios with virtually any classifier error, the research show that random sampling is comparable to (and possibly somewhat higher than) bagging but not as precise as boosting.

Application of KNN and Decision tree has been proved to provide higher accuracy in retail industry. Also, the study(Dietterich; 2000) states that boosting algorithm performs better compared to normal algorithms. To note, random forest and LGBM(Light Gradient Boosted Machine) has proved to be accurate but not applied much in retail domain. Hence, to conclude Decision tree, Random Forest, LGBM and KNN classifiers will be used to predict the consumer segment¿

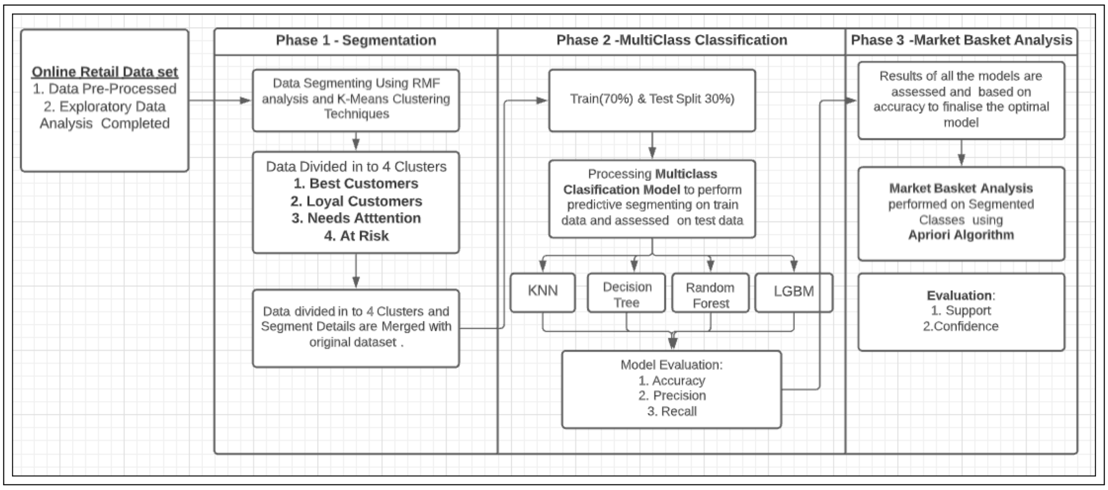
2.3 Market Basket Analysis

Christy et al. (2021) conducted a comparative study to identify an apt model for consumer segmentation and pointed out the need to conduct an experiment to apply association mining rules to the segmented customers, which is the motivation behind the second objective of this research. There are a number of algorithms such as SVM that help identify the association between independent variables, and have proven to be the appropriate model to build a recommendation system on high-dimentional data. The study illustrates how the parameterization was trained using precipitation data. There have been studies to develop an algorithm to recognize repeating item sets from a transaction dataset, and establishing association rules is one of the most commonly used data mining methodologies. Because of the intricacy, identifying common item-sets is difficult. Apriori first uses anti-monotonicity to search the database for frequent itemsets of size 1, adding up the counts for each item and collecting those that meet the minimal support condition that applies to any transaction data (Wu et al.; 2008).

Based on the examined literature, SVM, in contrast to Apriori, is a common data mining technique that produces better results in demographical and psychological data. Since this proposal is intended to be tested on transactional data, the Apriori algorithm is the best fit to recognize products that a customer will potentially be interested in.

3 Methodology

This research aims to implement a consumer segmentation and association mining technique on an online retail dataset, and a three-phased approach is used to attain this, as shown in Figure 1.The data mining approach follows CRISP-DM procedure. Firstly, the consumers are grouped into sets based on their shopping patterns. Recency, frequency, and monetary elements were considered to get a comprehensive knowledge of consumer behaviour. The RFM analysed dataset was then subjected to the K-Means clustering approach to determine the ideal number of segments necessary to classify the data such that each segment did not overlap with each other, resulting in four sections. Based on the approaches utilized, the four segments deemed appropriate for the dataset used in this study are 'Best Customers','Loyal Customers', 'Needs Attention' and 'At Risk'. This segmentation will help marketing team to tailor their marketing plan efficiently for respective consumer groups.



Online Retail Data set
1. Data Pre-Processed
2. Exploratory Data Analysis Completed
Phase 1 - Segmentation
Data Segmenting Using RMF analysis and K-Means Clustering Techniques
Data Divided in to 4 Clusters
1. Best Customers
2. Loyal Customers
3. Needs Atttention
4. At Risk
Data divided in to 4 Clusters and Segment Details are Merged with original dataset .
Phase 2 -MultiClass Classification
Train(70%) & Test Split 30%)
Processing Multiclass Clasification Model to perform predictive segmenting on train data and assessed on test data
KNN
Decision Tree
Random Forest
LGBM
Model Evaluation:
1. Accuracy
2. Precision
3. Recall
Phase 3 -Market Basket Analysis
Results of all the models are assessed and based on accuracy to finalise the optimal model
Market Basket Analysis performed on Segmented Classes using Apriori Algorithm
Evaluation:
1. Support
2.Confidence

Figure 1: Three Phased Approach

It is important to emphasize that the main objective of this study is not to define consumer segments, but rather to be able to anticipate consumer classes and verify that each consumer class is appropriately predicted. The customer segments derived by **RFM analysis and K-Means clustering will be the dependent variable** for the multi-class classification model built in phase 2, meaning that we will predict the extracted consumer segment using the transactional information available in the dataset. Then,

Column Name	Data Type	Column Description
InvoiceNo	Nominal	Each transaction is allocated a 6-digit integral number. If this code begins with the letter 'c,' it represents a cancellation.
StockCode	Nominal	Each different product is allocated a 5-digit integral number.
Description	Nominal	Name of the product
Quantity	Numeric	The number of units of each product (item) sold in a single transaction.
Invoice Date	Date and Time	The day and time when a transaction was generated.
Price	Numeric	The unit price of the product is in sterling (£).
CustomerID	Nominal	Integral number that is unique to each client.
Country	Nominal	The name of the country where a customer resides.

Table 1: Dataset Description

Market Basket Analysis (MBA) or Association Mining Rule will be applied to the segmented classes rather than the training corpus, which will help understand the prominent products in each category.

Secondly, a predictive model is developed that utilizes the transaction information accessible in the dataset as the independent variable, with the dependent categorical variable holding the segment classifier for each group, as previously mentioned. As a result, this model demonstrates a multi-class classification problem. Following the review of the literatures, **four predictive modelling algorithms** will be used to develop the model, and each model will be assessed on comparable grounds in order to understand and decide the final optimal model.

In the last phase, **the Association mining rule will be applied separately to each classed segment** to comprehend the products relevant to each group. Because the recency, frequency, and monetary value of each customer class varies, there may be a relevance in the product that each group prefers, which may be detected using association mining methods and this approach is utilized to develop a relevant product recommendation system for each consumer class.

3.1 Data Description

This study makes use of transactional data from a non-store online gift business situated in the United Kingdom (Online Retail II UCI)[1]. The firm provides a variety of unique giftware, and the dataset contains all transactional information from 1/12/2009 to 09/12/2011. Another piece of information is that the majority of clients are also wholesale sellers. **The dataset has a total 8 variables divided across 5 categorical and 3 numerical values and the total observations recorded is 1,067,371.**The column descriptors, their respective data type, and column names are all outlined in table 1.

[1] https://www.kaggle.com/mashlyn/online-retail-ii-uci

3.2 Data Cleaning

The initial analysis of the dataset using descriptive stats, info functions, and pandas profiling in Python assists in discovering the missing values, duplicate records, cardinality, skewness in the data, etc. A few notable inconsistencies discovered were as follows:

1. In the price column, there are 6202 records with zero values, which seems odd, and 5 rows with large negative values.
2. The orders originated from 43 distinct nations, but 91.9% of the orders came from the United Kingdom.
3. The Quantity field includes a negative minimum value, which looks to be erroneous since an order cannot be placed for a negative count. The field requires additional exploration.
4. The Customer ID and Description columns contain missing values. The **Customer Id** column has the **highest missing value of 22.8%** (i.e., 243007) and the data type is float.
5. There are **34335 duplicate** rows (i.e. 3.2%)
6. The distribution of quantity and price columns is skewed, indicating the presence of outliers in the columns.

The **assumption** made for this dataset were; unique customer ids were allocated to each customer; the Invoice number generated for each transaction was unique and the stock code for each product was unique. The duplicate entries were eliminated and since the missing records were in Customer Id and Description columns, the missing values were not discarded, instead they were replaced with 0 and ”NoValue” respectively for analysis purposes. The entries with negative values in the ’Price’ column were removed since it was not connected to any Customer id and were labelled as ’Bad Debts’. There were 22496 entries where the quantity was negative and the descriptions of 19863 records are missing for the given segment. In the given dataset 22.8% of the transaction doesn’t have customer id, 1.8% of the records are cancelled transaction and 75.5% of the transactions were successfully processed (refer Figure 2).

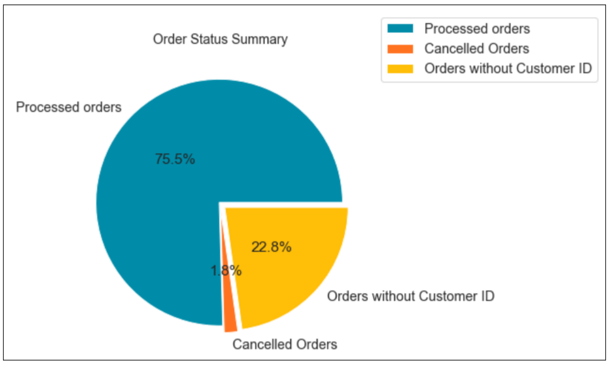


Figure 2: Order Status Summary Split

To assess such records and to interpret the data better, the dataset is partitioned into three unique data frames; one for cancelled transaction, another data frame where customer ID is 0, and the last data frame for the orders that were successfully processed. Also, by examining the country column it was identified that 91.9% of the order's revenue were being produced from the United Kingdom, the remaining 8.1% (refer Figure 3)would not bring much value to the analysis and thus may be discarded.

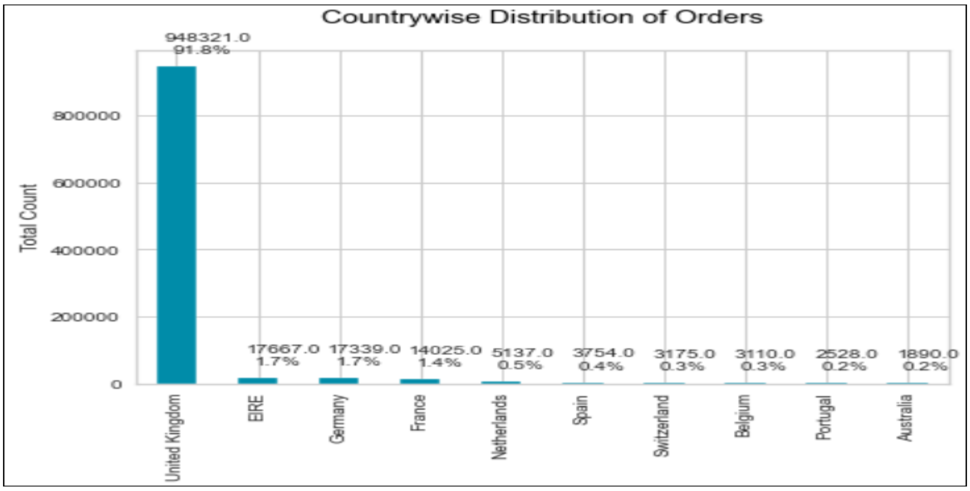


Figure 3: Countrywise split of Orders

Feature engineering(Zheng and Casari; 2018) was split across Data preparation and Implementation phase. In Data preparation phase, only **feature creation** is undertaken. **Feature Extraction and Feature Transformation** application is detailed in depth in the Implementation part (refer section 5)
This dataset includes the price and quantity of products purchased columns, using which the total billing amount (i.e. price * quantity) could be calculated, which would aid in comprehending the total revenue generated. The InvoiceDate might be further separated into year and month columns that would assist in analysing the purchase pattern or revenue trend on an annual and monthly basis.

Post data cleaning and feature creation, the dataset has 9 columns and 10,33036 records.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (Tukey et al.; 1977) was undertaken in order to identify patterns with the use of descriptive statistics and visualizations such as pie chart, bar plot, line graphs.

By analysing the successfully processed transactions, the important insights acquired were that, for the year 2009, solely Decembers' records were included in the dataset (refer Figure 4)and maximum sales were recorded in the month of November 2011 followed by November 2010. To evaluate the pattern, it has been found that the highest sales are recorded in the months of November, October, and September, and the least sales are recorded in the month of February. Further Analysis was undertaken based on orders received on an annual basis between the year 2010 and 2011. From the bar plot (refer Figure 4) it is also evident that the number of orders has reduced from 2010 to 2011, since the online shopping trend was blooming the reduction in sales could possibly indicate that the either there was a newly established competition in the market or the business did lose their revenue generating customers. Cancelled transactions were also studied, and it was revealed that two consumers had a maximum cancellation rate of 75%. The description for most of the cancelled transactions was indicated as "Manual", implying that most of the transactions cancelled were entered manually, and no product details were presented, which leads to inadequate data. The second most frequently cancelled item is the 'Regency Cakestand 3 tier'. As the reason is not given, there might be numerous causes, such as product damage, or delivery timeframes being high. There are no adequate details to come up with a remedial approach to reduce the product cancellation.

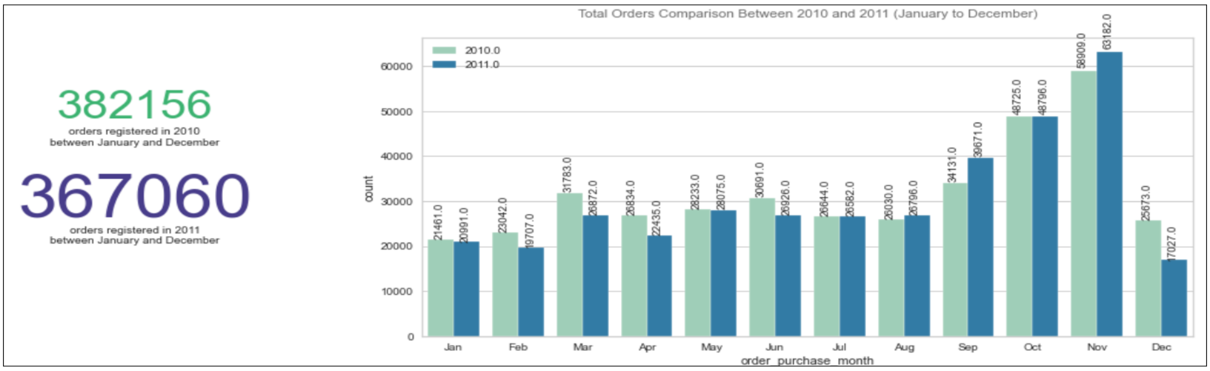


Figure 4: Online Retail Monthly Sales from 2010 to 2011

3.4 Customer Segmentation Techniques

This research uses RFM analysis with a k-Means clustering approach to identify the optimum number of consumer segments. This study blends RFM analysis with a k-Means clustering technique to discover the optimal number of customer segments.

RFM analysis: RFM analysis is a powerful data mining approach that is widely used to assess customers based on their previous purchasing history. This technique categorizes consumers based on three criteria: recency (R), frequency (F), and monetary value (M) (Christy et al.; 2021). **Recency** is used to identify when a customer last made a transaction. A lower number of recency indicates that the customer accesses the site on a regular basis. **Frequency**is defined as the total number of purchases made by a customer at a given time. The higher the value of frequency, the more loyal the firm's clientele are. **Monetary** is defined as the amount of money spent by a customer over a certain time period.

Each customer is assigned three unique rankings based on recency, frequency, and monetary factors. On data, the Quintile technique is used, and scoring is done on a scale of 1 to 4. The top quintile receives a score of 1, while the other quintiles get scores of 2, 3, and 4. The consumer categories may be evaluated using the RFM Scale(refer table 2).It is reasonable to assume that the scores have separate features. The RFM Score is calculated by summing the scores for recency, frequency, and monetary value.

K-Means Clustering: K-Means is a typical approach that takes as inputs the variables and the number of clusters and divides the data into the given number of classes with a high intra-class similarity(Likas et al.; 2003). K-Means is an iterative method that

Scale	Characteristics
1	Potential
2	Promising
3	Can't Lose Them
4	At Risk

Table 2: RFM Scale Description

computes the value of cluster centres before each iteration. To eliminate skewness in the data, recency, frequency, and monetary amounts are normalized using a conventional scaler approach. The scaled data is then subjected to the clustering process.

In addition, the Elbow Method[2] and the snake plot visualization were used to determine the optimal number of clusters.

3.5 Multiclass Classification Technique

Consumers are predicted to be in their appropriate clusters using four multiclass classification techniques: K-Nearest Neighbour (KNN), Decision Tree (DT), Random Forest (RF), and Light Gradient Boosting Model (LGBM). Despite their capacity to produce outstanding outcomes and requiring less time for implementation, Random Forest and LGBM are underutilized in the retail industry.The use of ensemble approaches is motivated by the fact that a group of models with comparable training results may have varied generalization abilities. Additionally, combining the results of various models minimizes the risk of picking a model that performs inadequately. Each of the strategies evaluated in this research is briefly described in the following sections.

1. **K-Nearest Neighbour (KNN):** In its most basic version, the KNN technique implies that all instances correspond to points in the n-dimensional space [3]. The standard Euclidean distance is often used to find an instance's nearest neighbours. In this scenario, the model was run using the default settings, and no parameters were optimized and the resulting output was used to evaluate the test data based also on default parameters.

2. **Decision Trees Classifier (DTC):** Decision trees are categorization techniques based on the divide-and-conquer principle (Quinlan; 1986). This implies that the original data set is gradually split into smaller subgroups based on the values of an explanatory variable selected according to an attribute selection criteria.In this scenario, the "criterion" and "max_depth" were optimized as "gini" and "2" respectively to train the model and the resulting output was used to evaluate the test data based on the trained parameters.

3. **Random Forest Classifier (RFC):** Random forests (Breiman; 2001) fall under the umbrella of ensemble approaches, since they mix numerous decision trees to exceed the performance of each individual decision tree. Typically, this approach allows for the resolution of issues caused by overfitting and distortion in data that would be difficult to resolve with a single tree. The random forest method generates each decision tree based on separate training sets that are drawn individually with

[2] https://vitalflux.com/k-means-elbow-point-method-sse-inertia-plot-python/

[3] https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote02_kNN.html

replacement from the training set. In this scenario, the "criterion","n-estimators" and "depths" were optimized as "gini", "100" and "10" respectively to train the model and the resulting output was used to evaluate the test data based on the trained parameters.

4. **Light Gradient Boosted Machine(LGBM) :** LightGBM, also known as the parallel voting DT technique, uses a histogram-based strategy to expedite training, minimize memory use, and combine advanced network connections to maximize parallel learning. Furthermore, LightGBM develops trees leaf-by-leaf, seeking for the leaf with the highest rise in variance to split (Ke et al.; 2017). In this scenario, the "n-estimators" and "learning_rate" were optimized as "1100" and "0.2" respectively to train the model and the resulting output was used to evaluate the test data based on the trained parameters.

5. **Apriori Algorithm** Apriori uses a "bottom-up" technique, in which regular subsets are expanded one item at a time, and sets of choices are assessed against the data. When no further successful expansions are detected, the operation comes to an end. The Apriori technique efficiently counts candidate items by combining breadth-first search with a hash tree structure. It generates candidate item sets of length k from item sets of length k-1 and discards those with an unusual sub pattern. It is often used to create a recommendation system and has proven to be beneficial for transactional datasets(Agrawal et al.; 1994). The support threshold used in this context was 0.03, giving a list of recommended items having support equal to or greater than the threshold.

3.6 Evaluation Metrics

1. **Multi-Class Classification Evaluation:**
   Accuracy, Confusion Matrix, Precision, Recall will be used to assess the multi-class classification model.

   - **Confusion Matrix:** It is an efficiency statistic for machine learning classification tasks that include two or more output classes. In terms of this study approach, an example of assessment for each category is provided below:
     True Positives: Actual loyal customers, Predicted loyal customers
     False Positives: Not a loyal customer, Predicted to be a loyal customer.
     False Negatives: A loyal customer, Predicted to be not a loyal customer.
     True Negative: Not a loyal customer, Predicted as not a loyal customer
   - **Accuracy:** Accuracy is one measure for assessing classification models. It is defined as the ratio of total correctly predicted value to the total values.
   - **Precision** Precision refers to how precise/accurate your model is in terms of how many of those predicted positives actually occur. When the costs of False Positives are high, accuracy is an appropriate statistic to utilize. For example, if the loyal or top customers are not accurately categorized, the consumers may miss out on the best deals or discounts, resulting in a loss. Precision could assist in evaluating the model from the aforementioned standpoint.

- **Recall** Recall measures the number of actual positives captured by the model by classifying them as positive (True Positive). In this study, if a client falls into the "at-risk" category but is misclassified as a "loyal customer," the company will be unable to take the necessary action on time, costing them customers, which directly ties to revenue. Hence, recall will be used as a yardstick to measure the performance of the multiclass classification model.

Also, the multiclass classification models are assessed using a **K-Fold validation** to estimate the competence of the model on new data.

2. **Association Mining Evaluation Matrix:** To analyse and evaluate the formation of an association model, a minimum degree of support and confidence is required (in this case, support should be greater than 0.03). Significant indicators for judging the quality of the models created for association mining rules are support and confidence.

   - **Support:** A rule's support specifies how often the things within that rule occur together.
   $$support(A\ implies\ B) = P(A, B) \tag{1}$$
   - **Confidence:** The confidence of a rule indicates the possibility of both the antecedent and the subsequent occurring in the same transaction. Confidence is defined as the conditional probability of the consequent given the antecedent.

   $$Confidence(AimpliesB) = P(B/A),\ which\ is\ equal\ to\ P(A, B)/P(A) \tag{2}$$

4 Design Specification

The architectural model implemented in this research is a two layer method, as illustrated in figure 5. The model includes a brief outline of the approaches used, as well as the tools and technologies employed.

The dataset used in this research is an Online Retail II public dataset accessible on Kaggle, the information was acquired in a csv file. A detailed description of the architecture is presented below.

1. **Tier 1 : Database Layer** In the database layer, data is extracted, data preprocessing is completed, which comprises data cleaning operations such as resolving missing values, feature creation, de-duplication, and also exploratory data analysis.

2. **Tier 2 : Business Logic Layer** The business logic layer illustrates the flow of data coming in from the database layer. It proceeds on to the RFM analysis and K-Means Clustering stage, where the feature is extracted based on the recency, frequency, and monetary value of the customer. Later, the retrieved characteristics are normalized and standardized. After segmenting the consumers into their respective clusters, a multiclass classification model is used on the dataset to forecast the customer segmentation. Each model is assessed based on the evaluation metrics. In the last stage, the association mining rule is applied to the various clusters to determine the product or item list relevant to each customer group.

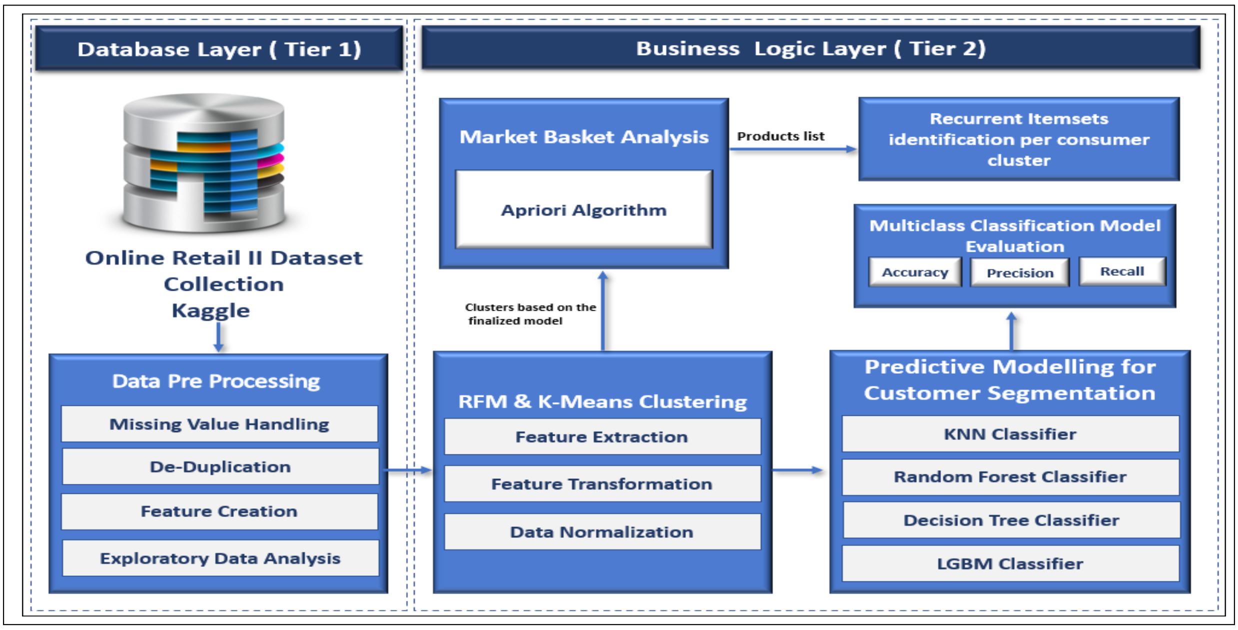


Figure 5: System Design Architecture

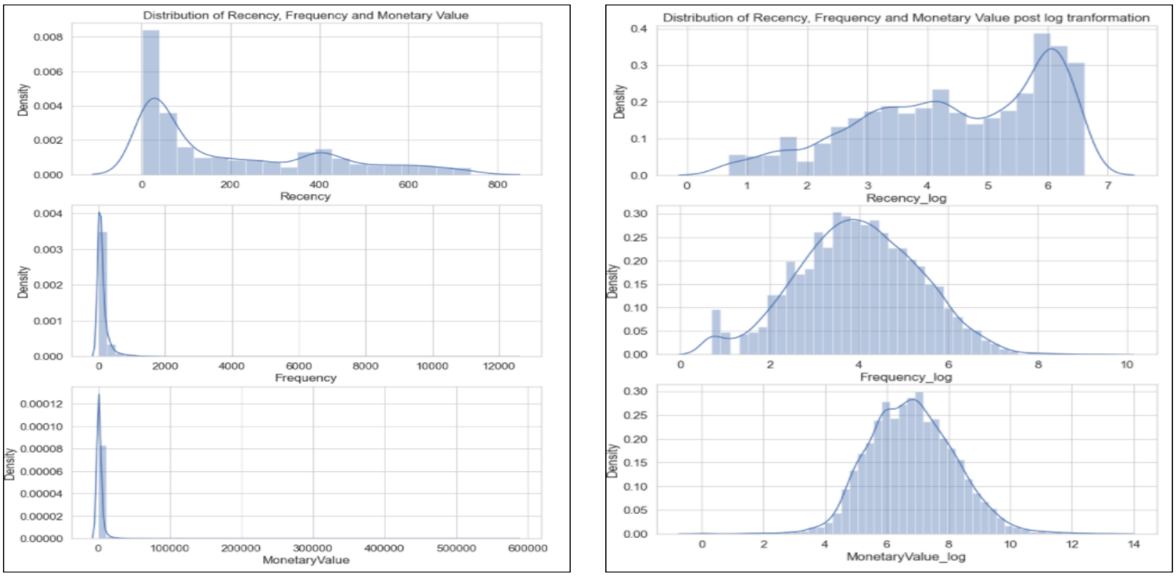
5 Implementation

This section discusses in detail the prerequisite data engineering performed to create customer segments, the hyperparameter tuning and class imbalance handling of data before implementing multi-class classification models, and implementation of Association Mining rule on segmented customers.

Tools and Technologies: Jupyter Notebook platform was used to for executing the code. The code was developed in python language using numerous python packages such as NumPy, Pandas, seaborn, Matplotlib, pandas profiling; etc were used to implement the code.

5.1 Feature Engineering

In this study, feature creation, feature transformation, and feature extraction were all applied as part of Feature Engineering. Features were created during the Data Preparation procedure (refer section 3.2). **Feature extraction and feature transformation** were conducted as part of the consumer segmentation approach, as detailed below.



(a) Distribution of R,F and M before transformation

(b) Distribution of R,F and M after transformation

Figure 6: Recency, Frequency and Monetary Value Distribution Pre- and Post-Data Transformation

This research makes use of transactional data, which contains information such as the transaction id, the number of orders placed, and the cost per unit. Using this data to comprehend each consumer's behaviour, one may readily determine the customer's recency, frequency, and total amount spent in the company. Each customer's **Recency, Frequency, and Monetary Value** characteristics were extracted using **RFM analysis**. Following feature extraction, the data was divided into four quartiles with values ranging from 1 to 4 based on the **RFM Score** characteristics (refer section 1). Four scores were assigned to the recency, frequency, and monetary value, respectively. For example, consumers who spend very little will be assigned a value of 4. The data was then **clustered** using the K-Means approach, which resulted in feature transformation.

Feature Transformation: Only when the data is regularly distributed and normalized will the K-Means approach be effective. Because the retrieved data for recency, frequency, and monetary value were not normally distributed, a log function was applied to the relevant feature to normalize the data Figure 6. After that, the data was scaled to standardize the data value. The **elbow method** (which employs the Sum square error (SSE) method or inertia) was used to determine the total number of segments over which the data should be divided. The idea is to find k (clusters) that still has a low SSE, and the elbow generally marks the point at which increasing k begins to have decreasing returns. We observe a reasonably smooth curve here, and it's unclear which number of k to choose, 3 or 4 or (refer Figure 7). The snake plot was used to identify the ideal number of clusters, and four clusters were chosen to be the best fit since, according to the **snake plot** inferences, four clusters all have a reasonable split and are not overlapping (referFigure 8). Based on recency, frequency and monetary mean on each cluster the consumers were divided into 4 clusters; cluster 0 = Best Customers, 1 = Needs Attention, 2 = At Risk and 3 = Loyal Customers.

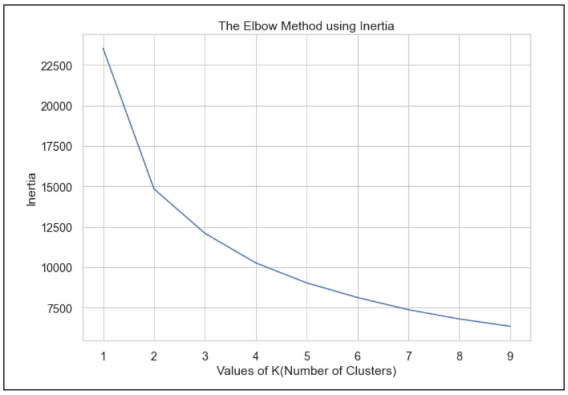
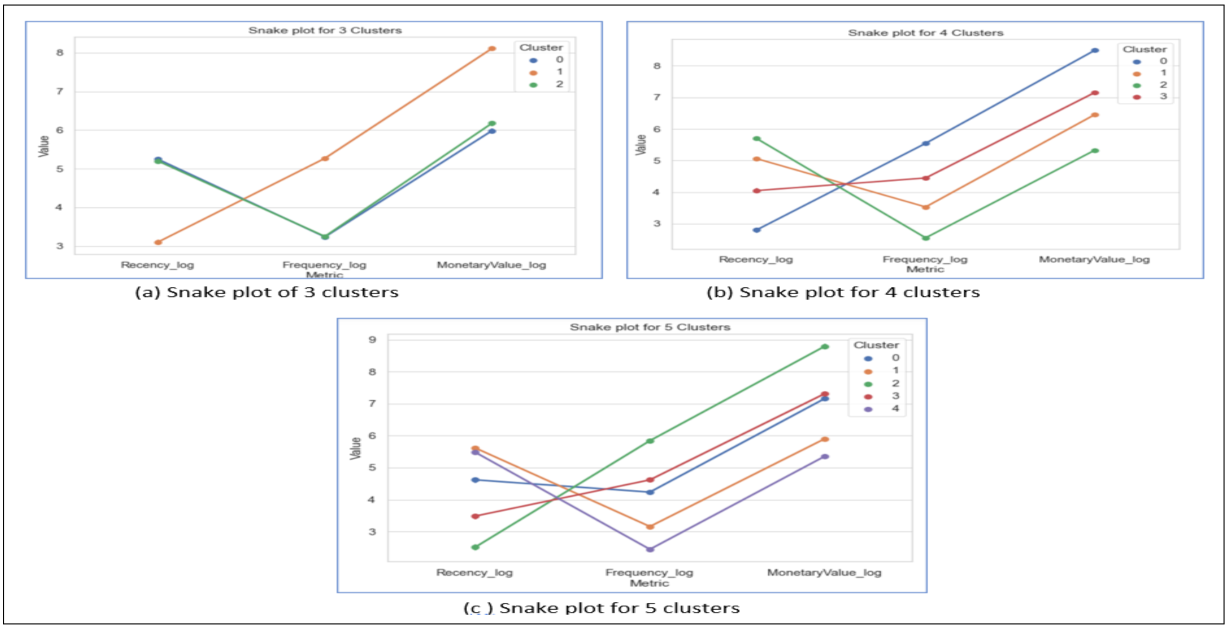


Figure 7: Elbow Method to Identify Optimal Number of Clusters



(a) Snake plot of 3 clusters

(b) Snake plot for 4 clusters

(c) Snake plot for 5 clusters

Figure 8: Snake plots for Clusters 3, 4 and 5 respectively

5.2 Data Imbalance

Post data segmentation, imbalance in cluster variable was checked by plotting a bar graph (refer Figure 9).

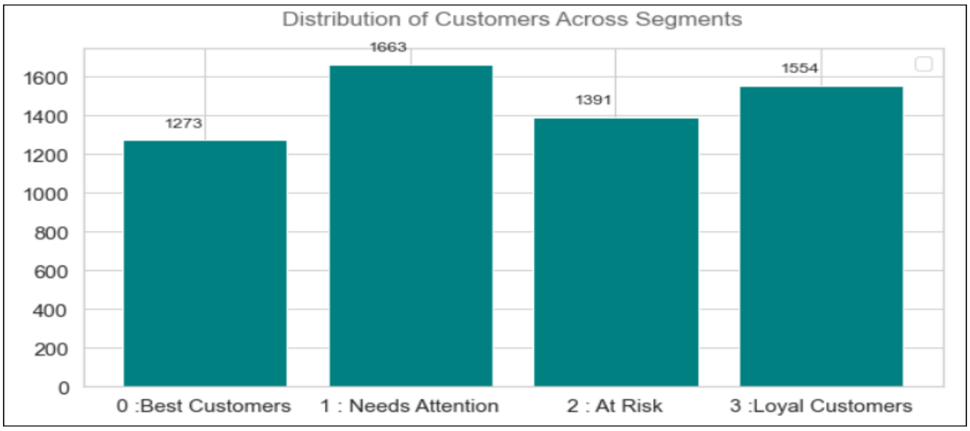


Figure 9: Distribution of Customers Across Segments

The bar plot clearly demonstrates that there is just a slight imbalance in the data split. Minor imbalances are seldom a concern, and therefore could not be treated as a conventional classification predictive modelling problem. A large class imbalance may be challenging to model and may require the use of specialized approaches. Because the data imbalance is so minor in this scenario, using an imbalance handling technique is not essential. However, **SMOTE (Synthetic Minority Oversampling Technique)** [4] is used to assess that a minor imbalance in the data is not causing any bias in the model. **A 70% - 30% stratified Train-Test split** was performed on the dataset to ensure that all the clusters are evenly distributed across the training and the testing data.

5.3 Classification Models and Hyperparameter Tuning

KNN, Decision Tree (DT), Random Forest (RF), and LGBM multiclass classification models were utilized to develop a prediction model that ensured clients were accurately allocated to their respective groups. **Hyperparameter tuning** was conducted on DT, RF and LGBM model to boost the performance of the model.

- The **KNN technique** application in retail industry has been proved effective in varied scenarios such as customer segmentation, customer retention,etc; (Abbasimehr and Shabani; 2021).In this scenario, the model was run using the default settings, and no parameters were optimized.

- The **Decision** trees are classification methods based on the divide-and-conquer notion. Each test result in the Decision Tree is indicated by an outgoing branch. In this research, the **"criterion" and "max_depth" were tuned as "gini" and 2** correspondingly to produce the best results.

[4] https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

- The **Random forests** (Breiman; 2001) are ensemble methods to data analysis. They combine numerous decision trees to exceed the performance of each individual decision tree by permitting rectification of challenges caused by overfitting and distortion in data that would be difficult to solve with a single tree. In this case, we **optimized the "depths" and "estimators" parameter by "10" and "100" respectively** to maximize the result.

- The **LightGBM**(Ke et al.; 2017), also termed as the parallel voting DT approach, is a histogram-based strategy to speed training. LightGBM builds trees leaf-by-leaf, aiming for the leaf with the biggest increase in variance to split. The parameters **"n_estimators" and "learning_rate" were tuned to "1100" and "0.2"** respectively to produce optimal output.

5.4 Association Mining Rule

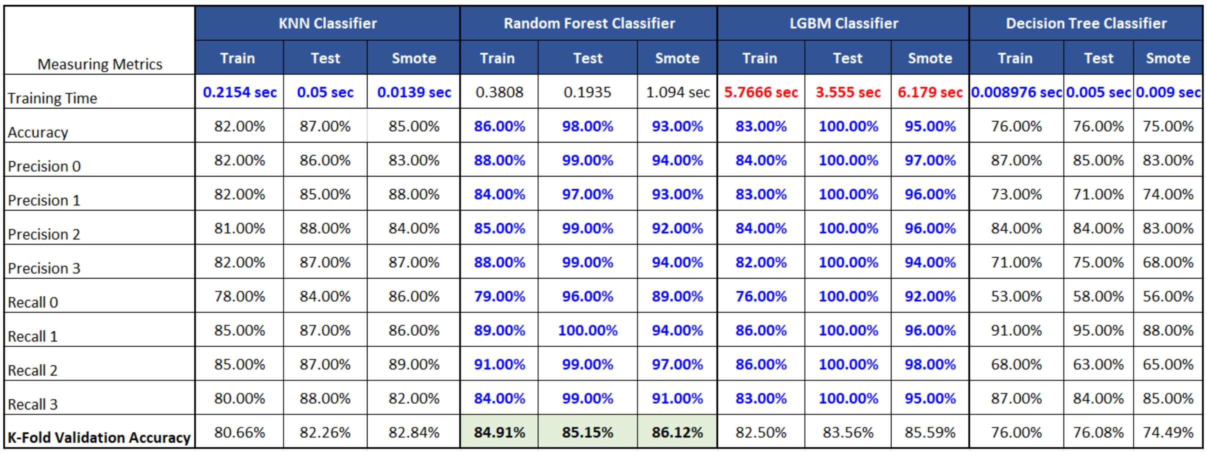
It is important to note that the second part of this study applies market basket analysis to clustered segments rather than the entire dataset. To achieve this, the dataset was partitioned into different datasets for each segment following cluster formation. Apriori algorithm is then applied to each unique datasets to determine the most frequent items bought by customers in their respective group. The product description attribute, which was of string data type, was encoded using **Label Encoder** technique before applying Apriori Algorithm. The frequent item sets were retrieved based on the two **support parameters being equal to or greater than 0.03 and 0.01** respectively.

6 Results and Discussions

Since this study had two aims, as indicated in (refer section introduction 1), each objective were evaluated differently.

6.1 Multiclass Classification Technique Performance

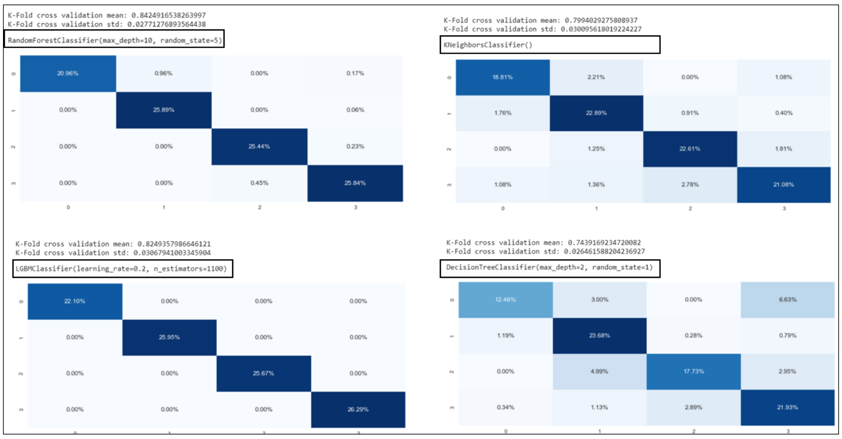
The classification models were evaluated based on the training time, and other metrics mentioned in the methodology (refer section 3.6). Figure 10 summarizes the results obtained.



Measuring Metrics	KNN Classifier			Random Forest Classifier			LGBM Classifier			Decision Tree Classifier		
	Train	Test	Smote	Train	Test	Smote	Train	Test	Smote	Train	Test	Smote
Training Time	0.2154 sec	0.05 sec	0.0139 sec	0.3808	0.1935	1.094 sec	5.7666 sec	3.555 sec	6.179 sec	0.008976 sec	0.005 sec	0.009 sec
Accuracy	82.00%	87.00%	85.00%	86.00%	98.00%	93.00%	83.00%	100.00%	95.00%	76.00%	76.00%	75.00%
Precision 0	82.00%	86.00%	83.00%	88.00%	99.00%	94.00%	84.00%	100.00%	97.00%	87.00%	85.00%	83.00%
Precision 1	82.00%	85.00%	88.00%	84.00%	97.00%	93.00%	83.00%	100.00%	96.00%	73.00%	71.00%	74.00%
Precision 2	81.00%	88.00%	84.00%	85.00%	99.00%	92.00%	84.00%	100.00%	96.00%	84.00%	84.00%	83.00%
Precision 3	82.00%	87.00%	87.00%	88.00%	99.00%	94.00%	82.00%	100.00%	94.00%	71.00%	75.00%	68.00%
Recall 0	78.00%	84.00%	86.00%	79.00%	96.00%	89.00%	76.00%	100.00%	92.00%	53.00%	58.00%	56.00%
Recall 1	85.00%	87.00%	86.00%	89.00%	100.00%	94.00%	86.00%	100.00%	96.00%	91.00%	95.00%	88.00%
Recall 2	85.00%	87.00%	89.00%	91.00%	99.00%	97.00%	86.00%	100.00%	98.00%	68.00%	63.00%	65.00%
Recall 3	80.00%	88.00%	82.00%	84.00%	99.00%	91.00%	83.00%	100.00%	95.00%	87.00%	84.00%	85.00%
K-Fold Validation Accuracy	80.66%	82.26%	82.84%	**84.91%**	**85.15%**	**86.12%**	82.50%	83.56%	85.59%	76.00%	76.08%	74.49%

Figure 10: Multiclass Classification Model Performance

Based on an examination of 10, we can deduce that the training time of all the models, except for the LGBM classifier, is relatively efficient. The models accuracy, precision, and recall ratings ranged from 76% to 100%. When compared to other models, the decision tree classifier achieves the lowest accuracy, obtaining 76 percent on the test data. The top two models in terms of the **least training time** were the **Random Forest Classifier (RFC) and the KNN classifier**. The figures seem to suggest that, in terms of test data accuracy, precision, and recall context, LGBM surpasses other classifier models and RF classifier is the second best (refer Figure 11) for confusion matrix. However, when examining performance based on k-fold validation, **RFC outperforms other models with a maximum accuracy of 85.15%. This study corroborates the findings of Zaklouta et al. (2011),** which reveal that random forests acquire state-of-the-art performance in several multi-class classification applications.



```
K-Fold cross validation mean: 0.8424916538263997
K-Fold cross validation std: 0.02771276893564438
RandomForestClassifier(max_depth=10, random_state=5)
```

```
K-Fold cross validation mean: 0.7994029275808937
K-Fold cross validation std: 0.030095618019224227
KNeighborsClassifier()
```

```
K-Fold cross validation mean: 0.8249357986646121
K-Fold cross validation std: 0.03067941003345904
LGBMClassifier(learning_rate=0.2, n_estimators=1100)
```

```
K-Fold cross validation mean: 0.7439169234720082
K-Fold cross validation std: 0.026461588204236927
DecisionTreeClassifier(max_depth=2, random_state=1)
```

Figure 11: Confusion Matrix and k-fold accuracy of Random Forest, KNN classifier, LGBM Classifier and Decision Tree Classifier

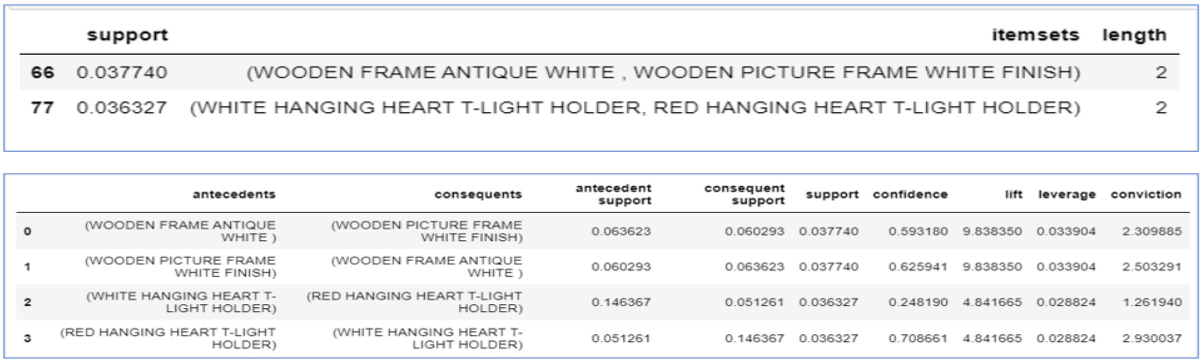
6.2 Market Basket Analysis Evaluation

To reiterate, the approach as mentioned in 1, the second objective of this research was to perform Market Basket Analysis on the segmented customers set rather than on the entire dataset. Apriori algorithm was applied to each set and was evaluated based on the support and confidence. The frequent item sets generated were evaluated for each consumer segment.

Segment 0 - Best Customers : As seen in figure 12, **4 frequent itemsets were identified** for Cluster 0, and they are wooden frame antique white, wooden picture frame white finish, and white hanging heart T-light holder and red hanging heart T-light holder, with a support of 0.037740 and 0.036327, respectively.

Segment 1 - Needs Attention and Segment 2 - At Risk : No frequent item sets were detected for the clients belonging to the aforementioned segmented clusters, even after decreasing the support parameter or reducing it to zero. The key reason behind no items being related was most of the customers in these categories only bought one item at a time, hence association mining rule didn't function on the mentioned consumer groups.

Segment 3 - Loyal Customer : As seen in figure 13, just 1 frequent itemsets were

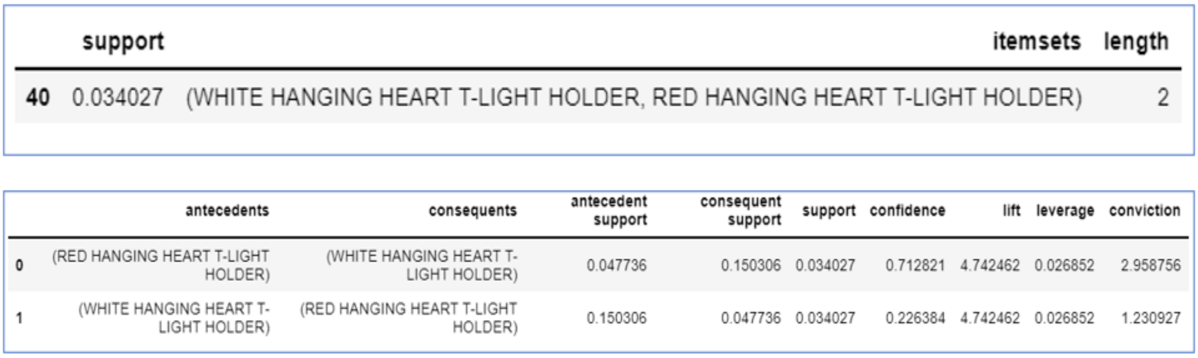


	support	itemsets	length
66	0.037740	(WOODEN FRAME ANTIQUE WHITE , WOODEN PICTURE FRAME WHITE FINISH)	2
77	0.036327	(WHITE HANGING HEART T-LIGHT HOLDER, RED HANGING HEART T-LIGHT HOLDER)	2

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(WOODEN FRAME ANTIQUE WHITE)	(WOODEN PICTURE FRAME WHITE FINISH)	0.063623	0.060293	0.037740	0.593180	9.838350	0.033904	2.309885
1	(WOODEN PICTURE FRAME WHITE FINISH)	(WOODEN FRAME ANTIQUE WHITE)	0.060293	0.063623	0.037740	0.625941	9.838350	0.033904	2.503291
2	(WHITE HANGING HEART T-LIGHT HOLDER)	(RED HANGING HEART T-LIGHT HOLDER)	0.146367	0.051261	0.036327	0.248190	4.841665	0.028824	1.261940
3	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.051261	0.146367	0.036327	0.708661	4.841665	0.028824	2.930037

Figure 12: Association Mining on Cluster 0: 'Best Customers

identified for Cluster 3, white hanging heart T-light holder and red hanging heart t-light holder, with a support of 0.034027.



	support	itemsets	length
40	0.034027	(WHITE HANGING HEART T-LIGHT HOLDER, RED HANGING HEART T-LIGHT HOLDER)	2

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.047736	0.150306	0.034027	0.712821	4.742462	0.026852	2.958756
1	(WHITE HANGING HEART T-LIGHT HOLDER)	(RED HANGING HEART T-LIGHT HOLDER)	0.150306	0.047736	0.034027	0.226384	4.742462	0.026852	1.230927

Figure 13: Association Mining on Cluster 3: 'Loyal Customers

The evaluation was conducted separately for Multiclass Classification model and Market Basket Analysis. In Multiclass classification problem, the results are analysed based on the predicting capability of classification models across each class. The models that showcase a high True Positive and True negative rate against each class is deemed optimal for prediction. Based on the evaluation metrics, the LGBM classifier has attained the highest accuracy on test data. However, when **taking into account the K-fold validation result and training duration of the models, the Random Forest classifier provides the best result**. The K-fold validation score assists in anticipating the performance of the models on new data. Therefore, the **higher the K-fold validation score, the better the model**. Also, while taking into consideration the **training time**, **Random Forest** was able to train the model in **0.19 seconds** compared to the best model, **LGBM**, which took **3.555sec** to train the model, which is comparably quite high. Hence, we might claim that Random Forest is the optimal multiclass classification model from every perspective for forecasting consumer groups. **Limitations with the multi-class classification strategy were that before applying the model, the data should be aggregated by total orders and quantity for each customer id. Doing otherwise resulted in overfitting of the model and poor accuracy.**

The objective of the application of association mining rules was to identify basket item sets relevant for each consumer category. From this experiment, the association mining rule worked only on the best customer segments (clusters0) and loyal customers (cluster 3). However, it failed to retrieve any relevant itemsets for the consumer class that belonged to segments ”At Risk” (cluster 1) and ”Needs Attention” (cluster 2). **The rationale for not retrieving items was that the customers in clusters 1 and 2 only purchased one item each, and hence the algorithm was not able to establish any relationship between the items themselves, resulting in no recommended items**. Also, to compare the outputs generated for customer cluster 0 and cluster 3, the products that were displayed for cluster 3 (Loyal Customer) were also included in the products displayed for cluster 0 (Best Customers), proving that a holistic recommendation system would be a better approach as there is a high possibility that for some classes, the system fails to produce a list of suggested items and the items generated for one class may be a subset of items for another class.

So, within this dataset it has been identified that to provision recommendations using Apriori algorithm, either of the two prerequisites should be fulfilled.
1. There should be two or more items tagged against each customer,
2. The application of the Apriori algorithm should be performed on the holistic transaction data.

7 Conclusion and Future Work

This research provides a model, underpinned by data mining classification methods, that forecasts consumer segments based on the transactional information available from an online retail store. This research indicates that prediction modelling is effective in the retail sector, and that decision makers may use such models to competently construct marketing strategies and policies, as well as to develop a recommendation system.

Based on the proposed research, the objectives and the research question stated in section 1, **the results indicate that the feasibility of using a machine learning approach integrated with RFM analysis and K-Means clustering stands valid and produces optimal results.** However, the **approach proposed for the latter seem to have achieved a partial success and produced knowledge on the best practices to inculcate within the subject for better results.** According to the results, the random forest model was able to attain accuracy levels of about 98% on the test data and the maximum K-fold test validation accuracy of 85.15% with a minimum training time of 0.19 seconds in the context of the initial objective of the research and was inline with results of Zaklouta et al. (2011), who claimed that random forests provide the best performance n a variety of multi-class classification applications. Furthermore, the experiment conducted in this research suggests that, in this particular case study, the random forest classification strategy produces the best results, while the decision tree classifier produces the worst. Regarding the **second research question**, which is an extension of the previous work (Christy et al.; 2021).The **model provided partial success** because products were recommended for two classes and no products were recommended for the remaining two classes. The instance where the customer ordered only one item per invoice failed to generate product recommendations, **thus rejecting our hypothesis and favouring the alternate hypothesis**. It can be deduced that devising a recommendation system based on the Apriori algorithm would be best suited when applied to the entire transaction or behavioural dataset, or at least it’s when there are multiple items purchased by a customer in each group. **Limitations of this research were that the analysis was undertaken on transaction data; no customer demographic or psychological details were available. Application of the same approach im-**

plemented in this research could yield different result when applied on data when segmented using consumer demographical or psychological data.

As part of future study, the association mining rule on segmented customers can be examined on a dataset incorporating consumer demographics information, where consumers could also be classified based on demographical features and not solely on the transactional or behavioural data (which was used in this study).

Acknowledgement

I'd want to use this opportunity to express my sincere thanks and warm regards to Prof. Athanasios Staikopoulos, who has served as a guide and mentor to me during this project. His persistent support and guidance have helped me to deal with the challenges brought by my research.

References

Abbasimehr, H. and Shabani, M. (2021). A new framework for predicting customer behavior in terms of rfm by considering the temporal aspect based on time series techniques, *Journal of Ambient Intelligence and Humanized Computing* **12**(1): 515–531.

Agrawal, R., Srikant, R. et al. (1994). Fast algorithms for mining association rules, *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215, Citeseer, pp. 487–499.

Aguinis, H., Forcum, L. E. and Joo, H. (2013). Using market basket analysis in management research, **39**(7): 1799–1824. Publisher: SAGE Publications Inc.

Aly, M. (2005). Survey on multiclass classification methods.

Azuaje, F. (2006). Witten IH, frank e: Data mining: Practical machine learning tools and techniques 2nd edition: San francisco: Morgan kaufmann publishers; 2005:560. ISBN 0-12-088407-0, £34.99, **5**(1): 51, 1475–925X–5–51.
URL: *https://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-5-51*

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Camilleri, M. A. (2017). Market segmentation, targeting and positioning.

Chan, C. C. H. (2008). Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer, **34**(4): 2754–2762.

Cheng, C.-H. and Chen, Y.-S. (2009). Classifying the segmentation of customer value via RFM model and RS theory, **36**(3): 4176–4184.

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.

Christy, A., Umamakeswari, A., Priyatharsini, L. and Neyaa, A. (2018). Rfm ranking – an effective approach to customer segmentation, *Journal of King Saud University - Computer and Information Sciences* **33**.

Dietterich, T. (2000). Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, *Machine Learning* **40**(2): 139–157.

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.

Hiziroglu, A. (2013). Soft computing applications in customer segmentation: State-of-art review and critique, *Expert Systems with Applications* **40**: 6491–6507.

Hosseini, S. M. S., Maleki, A. and Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop crm methodology to assess the customer loyalty, *Expert Systems with Applications* **37**(7): 5259–5264.

Hu, Y.-H. and Yeh, T.-W. (2014). Discovering valuable frequent patterns based on rfm analysis without customer identification information, *Knowledge-Based Systems* **61**: 76–88.

Kaur, M. and Kang, S. (2016). Market basket analysis: Identify the changing trends of market data using association rule mining, **85**: 78–85.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree, *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds), *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc.

Kulkarni, R. (2012). Association rule-extracting knowledge using market basket analysis, **1**.

Likas, A., Vlassis, N. and Verbeek, J. J. (2003). The global k-means clustering algorithm, *Pattern Recognition* **36**(2): 451–461.

McCarty, J. A. and Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression, **60**(6): 656–662. Publisher: Elsevier.

Mishra, H., Shuchi and Tripathi, S. (2017). *A Comparative Study of Data Clustering Techniques*.

Namvar, M., Khakabimamaghani, S. and Gholamian, M. R. (2011). An approach to optimised customer segmentation and profiling using rfm, ltv, and demographic features, *International Journal of Electronic Customer Relationship Management* **5**(3–4): 220–235.

Ngai, E. W. T., Xiu, L. and Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification, *Expert Systems with Applications* **36**(2, Part 2): 2592–2602.

Olson, D. and Chae, B. (2012). Direct marketing decision support through predictive customer response modeling, *Decision Support Systems* **54**(1): 443–451.

Olson, D. L. and Delen, D. (2008). *Advanced Data Mining Techniques*, Springer Science Business Media.

Quinlan, J. R. (1986). Induction of decision trees, *Machine Learning* **1**(1): 81–106.

Raju, P. S., Bai, D. V. R. and Chaitanya, G. K. (2007). Data mining: Techniques for enhancing customer relationship management in banking and retail industries, **2**(1): 8.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O., Tiwari, A., Er, M., Ding, W. and Lin, C.-T. (2017). A review of clustering techniques and developments, **267**: 664–681.

Tsai, C. Y. and Chiu, C. C. (2004). A purchase-based market segmentation methodology, **27**(2): 265–276.

Tukey, J. W. et al. (1977). *Exploratory data analysis*, Vol. 2, Reading, Mass.

Verhoef, P. C. (2003). Understanding the effect of customer relationship management efforts on customer retention and customer share development, **67**(4): 30–45. Publisher: SAGE Publications Inc.

Wu, X., Kumar, V., Ross, Q., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z.-H., Steinbach, M., Hand, D. and Steinberg, D. (2008). Top 10 algorithms in data mining, **14**(1): 1–37.

Zaklouta, F., Stanciulescu, B. and Hamdoun, O. (2011). Traffic sign classification using k-d trees and random forests, *The 2011 International Joint Conference on Neural Networks*, p. 2151–2155.

Zheng, A. and Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*, O'Reilly Media, Inc. Google-Books-ID: sthSDwAAQBAJ.