****

IS424 - Data Mining & Business Analytics (G2T7)

**Predicting value segments (market segmentation) of E-Commerce consumers based on their profitability using Clustering and Classification Methods**

**Dr. Huang Zhiwu**

9th November 2022

| A S Subhashini | **01392746** |
| --- | --- |
| Amir Syafiq Bin Abu Saiere | **01370609** |
| Daniel Choi | **01393238** |
| Helisa Binte Salim | **01383873** |
| Ong Xin Ya, Jasmine | **01396585** |
| Jaye Ding | **01371285** |

[**1. Introduction**](#_ylkpqfeqe6zs) **3**

[1.2 Problem Statement](#_inuokxnait7u) 3

[**2. Motivation**](#_rw6943bd92tk) **3**

[**3. Literature Review**](#_3llwqpmh5fhs) **5**

[3.1 A machine learning framework for customer purchase prediction in the non-contractual setting](#_8ep10jvx9o62) 5

[3.1.1 Summary](#_13b9086rdq8s) 5

[3.1.2 Techniques](#_wzykcytw6a3i) 5

[3.1.3 Pros & Cons](#_viuwqjb4ajox) 5

[3.1.4 Data Used](#_mpdw6whdej48) 6

[3.1.5 Results](#_1nyp3u5e2wn7) 6

[3.1.6 Key Takeaways](#_8a0csshmrzox) 6

[3.2 An empirical study on Customer Segmentation by Purchase Behaviours using a RFM model and K-means algorithm](#_a7tbvf59wytt) 7

[3.2.1 Summary](#_34fl9zs3h3ly) 7

[3.2.2 Techniques](#_x82su32em30i) 7

[3.2.3 Pros & Cons](#_y3bkn7kr0492) 7

[3.2.4 Data Used](#_qrdwiwdtydvl) 8

[3.2.5 Results](#_kvie7x281gk2) 8

[3.2.6 Key Takeaways](#_s4bhclff2f38) 8

[3.3 Article 3: Data accuracy’s impact on segmentation performance: Benchmarking RFM analysis, logistic regression and decision trees](#_svpbir1fkuh5) 8

[3.3.1 Summary](#_cimij21xyxyz) 8

[3.3.3 Pros & Cons](#_toya1r855zir) 9

[3.3.4 Data Used](#_u16yl8eiilf6) 9

[3.3.5 Results](#_4mjomxkm2elg) 9

[3.3.6 Key Takeaways](#_z99xhdbj1tzg) 10

[**4. Dataset**](#_yddvr8d37ibd) **10**

[**5. Methodology**](#_k4ybrod50mzt) **12**

[5.1 Data Preprocessing](#_uzqk69qfrpqz) 13

[5.1.1 Merging Data and Preliminary Table Selection](#_ykajva2h9wkn) 13

[5.1.2 Exploratory Data Analysis and Preliminary Feature Selection](#_lx6gjfcaolmx) 14

[5.1.3 Data Cleaning](#_er67bgztflk) 18

[5.2 Data Transformation](#_tm73rrj6v6jy) 18

[5.2.2 Feature Engineering (RFM)](#_xf14xf2246h4) 18

[5.2.3 Normalisation](#_ji1ydrglbid3) 19

[5.3 Data Reduction](#_lbyptku3phvh) 20

[5.3.1 Principal Component Analysis (PCA)](#_flrsg3fhitsi) 20

[5.4 Customer Segmentation](#_2shplw4rm0nu) 21

[5.4.1 RFM Scoring](#_6waw5t47bwll) 22

[5.4.2 K Means Clustering based on continuous variables](#_u8y4kqf8vel7) 27

[5.4.3 K-Prototype Clustering based on both categorical and numerical variables](#_o6q7irtj1met) 30

[5.4.4 Cluster/Segmentation Evaluation](#_3717wsafg7qt) 32

[5.4.5 Final Clusters after chosen K-Means clustering method](#_oqc4yjf7fr6w) 34

[5.5 Classification Modelling](#_tiruv1d1tl6x) 34

[5.5.1 Decision Tree](#_2w58mpakjcrt) 35

[5.5.2 Random Forest](#_25ii6hwxxyjx) 37

[5.5.3 XGBoost](#_pyzgeze3jq1) 37

[5.5.4 Classification Model Results & Discussion](#_4dv4be3leaqc) 38

[**7. Conclusion & Future Work**](#_4p0k1b135ob0) **38**

[7.1 Conclusion](#_2zr1udyrkpb3) 38

[7.2 Future Work](#_t19sj76fof8b) 39

[7.2.1 Businesses / Marketers](#_wacwlf1tj005) 39

[7.2.2 Academia / Research](#_l4ny3u5jzipe) 39

[**8. References**](#_waa5tmdnvdmj) **40**

# 

# 1. Introduction

The digital economy in Southeast Asia is expected to reach $1T in gross merchandise value (GMV) by 2030 and is on course to reach $200B GMV by end 2022 (Bain et al., 2022). The leading sectors in the digital economy include e-commerce, financial services, deliveries (food and transport), and online media. The pandemic was a catalyst for the adoption and growth of these digital platforms. E-commerce, specifically, experienced rapid growth with increased frequency, time and money spent on these platforms as consumers stayed at home. However, e-commerce’s steep acceleration has also brought about a slew of issues. Firstly, retailers have a larger, more diverse set of consumers now more than ever. Secondly, offering products on e-commerce is no longer optional to remain competitive which means an additional distribution channel to manage. Lastly, retailers have to navigate erratic demand and pricing driven by the promotional efforts of competing e-commerce platforms. Nevertheless, it has also provided retailers with access to vast amounts of consumer data that can be used to mitigate these additional challenges. The opportunity for growth in e-commerce lies in deeper engagement, including more frequent and valuable orders, subscriptions or cross-selling of other products and services such as financial services. Therefore, retailers that are able to fully harness consumer data to continuously identify levers for growth for different consumer segments would come out on top.

## 1.2 Problem Statement

The project, therefore, aims to construct a more dynamic & efficient process of segmentation via the combined use of clustering and classification algorithms, enabling firms to predict the value segments of their consumers more responsively.

# **2. Motivation**

Firstly, effective customer management and resource allocation requires corporations to identify the most optimal allocation of resources to marketing and sales. The ability to segment consumers based on their purchasing behaviour and profitability will allow retailers to prioritise their marketing and sales resources on re-targeting and servicing their most profitable customers, ensuring that they retain this segment. This would also allow the company to de-prioritize or improve the profitability of less profitable consumer segments. Additionally, understanding the characteristics of the firm’s most profitable and least profitable segments will also assist in shaping the firm’s customer acquisition strategy, further paving the way for the firm’s path to increased profitability. Therefore, an effective method of segmenting consumers based on their profitability is well worth examining.

|  |  |
| --- | --- |
| Figure 1: Current process of consumer segmentation | Figure 2: Suggested process of consumer  segmentation |

Secondly, many firms today employ suboptimal methods of segmenting their consumers. Figure 1 lays out a common process undertaken by marketing teams across multiple firms and industries (Refer to Figure 5 for Real World Example). Through this method, firms segment their consumers using business models such as the Customer Loyalty Model (Dick & Basu, 1994) or the Apostle Loyalty Model (Jones & Sasser, 1995). While this works at a broader level, such models do not label consumers at an individual level, preventing companies from leveraging consumer data to provide more targeted and personalised marketing messages such as through Direct Email Marketing.

For more data-savvy companies, simple clustering algorithms like K-Means are used. After the consumer segments are defined, they are used to craft marketing or business strategies. One major drawback of this process is that it is static. The segmentation is done only once and to classify a new customer or to reclassify an existing customer, the complete clustering model will have to be run again, which is inefficient. As firms strive to provide more targeted and personalised marketing, a more dynamic process of classifying and reclassifying consumers is required.

As such, our team proposes a new process of segmenting consumers that combines the strengths of using business domain knowledge as well as clustering and classification algorithms (Figure 2). In the suggested process, the firm would first segment its consumers using a clustering algorithm. A classification model is then built based on the defined consumer segments or clusters. After which, new customers can be classified based on the predefined segments. Business domain knowledge remains essential in feature selection, engineering and or selecting the appropriate number of segments.

This process allows the business to dynamically assign a segment to new customers without running the entire clustering model again, increasing responsiveness and efficiency. The business can choose to re-run the clustering algorithm when there is sufficient change in their consumer base or after sufficient time has lapsed. Thus, the suggested segmentation process improves the existing process while allowing for firms to leverage modern machine learning algorithms without disposing their business expertise.

# **3. Literature Review**

## **3.1** **A machine learning framework for customer purchase prediction in the non-contractual setting**

### 3.1.1 Summary

Martinez et al. (2020) aimed to compare different machine learning algorithms to identify the algorithm that can best predict whether a customer is going to make a purchase within a certain time frame in the near future. The algorithms compared were Lasso Logistic Regression (LASSO), Gradient Tree Boosting, and Extreme Learning Machine (ELM). The researchers also proposed a new set of customer features that is derived from times and values of previous purchases. Using a data set containing more than 10 000 customers and a total number of 200 000 purchases the researchers obtained an accuracy score of 89% and an AUC value of 0.95 for predicting next moth purchases on the test data set.

### **3.1.2 Techniques**

The following algorithms were used in this study:

1. Lasso Logistic Regression
2. Gradient Tree Boosting
3. Extreme Learning Machine

### **3.1.3 Pros & Cons**

| **Pros** | **Cons** |
| --- | --- |
| Significant investment in feature engineering of time-series data and product categories to increase predictive capability of models. | Heavy feature engineering with 274 features using advanced time-series methods that may not be easily implemented for companies that do not have dedicated data scientists or statisticians. |
| One of few research papers predicting B2B sales in a non-contractual setting. | Requires large amounts of historical data with customers having made multiple purchases across long periods of time (6 years). Not as applicable to startups, companies that have only started collecting data or areas with infrequent purchases. |
| Strong literature review, building on past research. | Does not predict purchases value of future transactions. |

*Table 1: Pros & Cons of the techniques used in Martinez et al. (2020)*

### **3.1.4 Data Used**

This study used transactional data provided by a large manufacturer located in central Europe. The data have been gathered from transactions of the B2B unit, which have been recorded from January 2009 until May 2015. In this study, only transactions of customers whose first purchase was at least six months ago were considered due to insufficient information in the other cases.

The Data set contains 192,470 orders for all customers., where researchers took January 2009 as month A = 1 such that May 2015 corresponds to month 77. The time period for computing the feature values is taken as T = 24. The transactions belong to K = 10136 different customers from 125 different countries, and orders were aggregated on a monthly level.

### **3.1.5 Results**

The researchers obtained an accuracy score of **89%** and an AUC value of **0.95** for predicting next month purchases on the test data set using XGBoost.

### **3.1.6 Key Takeaways**

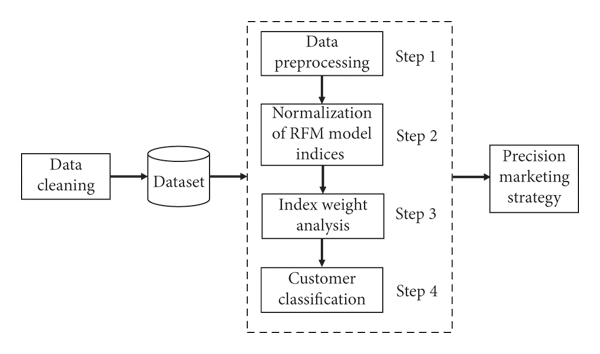
This paper offered great insight into the potential use of time-series data for customer purchase prediction. Unfortunately, our dataset does not allow us to utilise the same time series techniques utilised here due to a lack of large amounts of historical data and frequency of purchases as can be seen in subsequent sections. Nevertheless, we were also able to gain more insights on how to use XGBoost modelling for transactional data.

## **3.2 An empirical study on Customer Segmentation by Purchase Behaviours using a RFM model and K-means algorithm**

### 3.2.1 Summary

Wu et al. (2020) combined RFM (recency, frequency, and monetary) model and K-means clustering algorithm to conduct customer segmentation and value analysis of online sales data from Nov 2017 to Apr 2019 provided by a company in Beijing, China. The RFM method and K-Means Clustering algorithm was chosen based on existing literature such as Hughes, 1994 and Y. Jiang, D. Wu, Z. Deng et al, 2017.Customers were classified into four groups based on their RFM scores. Researchers then suggested different CRM (customer relationship management) strategies to increase customer satisfaction. Upon adoption of the strategies, the number of active customers increased by 529 while the total purchase volume increased by 279%, and the total consumption amount increased by 101.97%.

### **3.2.2 Techniques**

****

*Figure 3: An overview diagram of how the techniques being used*

The following techniques were used in this study:

1. Feature Engineering to create RFM variables
2. K-Means Clustering

### **3.2.3 Pros & Cons**

| **Pros** | **Cons** |
| --- | --- |
| Clear and defined scope of project built on existing literature. | Only considered RFM as variables and K-Means as a clustering algorithm. No effort done to compare different algorithms or utilise more variables to fit the specific use case. |
| Implemented marketing strategies based on cluster results to show effectiveness of segmentation . | No evaluations metrics used to evaluate clusters |
|  | Difficult to prove causal relation between segmentation results and actual increase in customers or improved KPIs. |

*Table 2: Pros & Cons of the techniques used in Wu et al. (2020)*

### **3.2.4 Data Used**

The study utilised private online transaction data collected from November 2017 to April 2019 from an online enterprise in Beijing, China. However, no information available on what variables were available in raw data.

### **3.2.5 Results**

No evaluation metrics used to evaluate clustering results. However, with the segments and marketing strategies for each cluster suggested, the number of active customers grew by 529. The total purchase volume increased by 279%, and the total consumption amount increased by 101.97%.

### **3.2.6 Key Takeaways**

This paper serves as a main reference and inspiration for our segmentation approach as explained in subsequent sections. Our project will seek to explore the use of RFM and K-Means clustering to segment consumers and explore methods to improve on this paper by utilising variables beyond RFM as well as to consider categorical variables.

## **3.3 Article 3: Data accuracy’s impact on segmentation performance: Benchmarking RFM analysis, logistic regression and decision trees**

### **3.3.1 Summary**

Coussement et al. (2014) aimed to investigate the influence of problems with data accuracy – an important dimension of data quality – on three prominent segmentation techniques for direct marketing: RFM (recency, frequency, and monetary value) analysis, logistic regression, and decision trees. The study recommends the use of decision trees in the context of customer segmentation for direct marketing, even under the suspicion of data accuracy problems

**3.3.2 Techniques**

The techniques used in this study are in the following:

1. CHAID Decision Trees
2. Direct RFM Segmentation
3. Logistic Regression

### **3.3.3 Pros & Cons**

| **Pros** | **Cons** |
| --- | --- |
| Implemented a fivefold cross-validation. | Only RFM variables were used for comparison |
| Robust experimental design for algorithm comparison. |  |
| Strong implications for marketers. |  |

*Table 3: Pros & Cons of the techniques used in Coussement et al. (2014)*

### **3.3.4 Data Used**

Two empirical direct marketing data sets provided by the Direct Marketing Educational Foundation (DMEF). The first data set originates from a multi-division mail-order catalogue business and contains information on 96,551 customers of which 2.46% responded to the mailing. The second data set contains 99,200 members of a non-profit organisation representing a response rate of 27.43%. Therefore, these two data sets allow for the ability to test the sensitivity of the three segmentation techniques to data accuracy problems at varying levels of response (low: 2.46% versus high: 27.43%).

### **3.3.5 Results**

The results demonstrate that (1) under optimal data accuracy, decision trees are preferred over RFM analysis and logistic regression; (2) the introduction of data accuracy problems deteriorates the performance of all three segmentation techniques; and (3) as data becomes less accurate, decision trees retain superior to logistic regression and RFM analysis.

### **3.3.6 Key Takeaways**

This literature review gave our group some potential directions we could take on using RFM analysis with data mining methods such as decision trees, once again giving us the assurance on the viability of such methods because of past research on such topics. The results of the study also gave us further considerations on evaluation on whether Decision trees analysis is better than RFM analysis - an area our group wanted to explore further into our project.

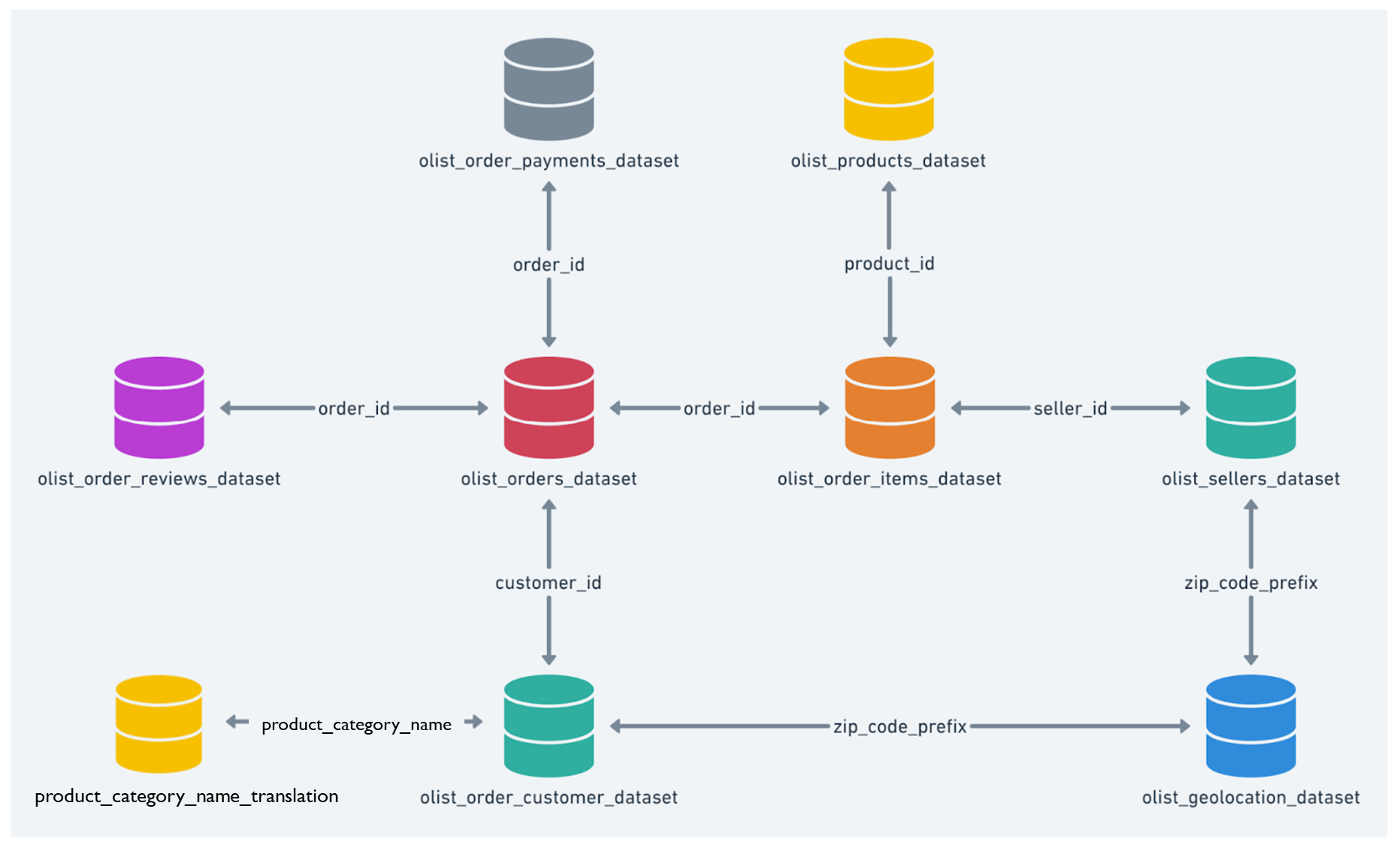
# **4. Dataset**

The dataset was originally extracted from Kaggle: Brazilian E-Commerce Public Dataset by Olist . Olist is the largest department store in the Brazilian marketplace and the dataset has information of 100,000 orders made at multiple marketplaces in Brazil from 2016 to 2018 (Olist & Sionek, n.d.). We chose this dataset because it is real world e-commerce data. Furthermore, datasets used by the Literature Reviews that we have researched on were private and unavailable.

The table below gives an overview of of each dataset:

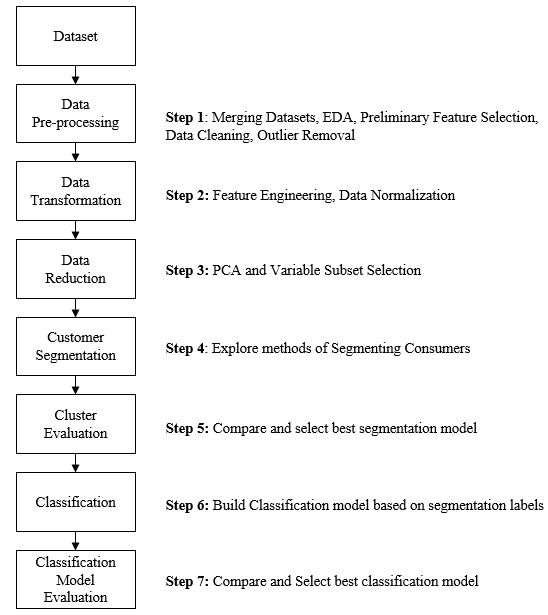
| **Dataset** | **Description** |
| --- | --- |
| Customer  (olist\_customers\_dataset) | This dataset has information about the customer and its location. It is used to identify unique customers in the orders dataset and to find the orders delivery location. |
| Geolocation  (olist\_geolocation\_dataset) | This dataset has information of Brazilian zip codes and its latitude and longitude coordinates. It is used to plot maps and find distances between sellers and customers. |
| Order Items  (olist\_order\_items\_dataset) | This dataset includes data about the items purchased within each order. |
| Order Payments (olist\_order\_payments\_dataset) | This dataset includes data about the orders payment options. |
| Order Reviews  (olist\_order\_reviews) | This dataset includes data about the reviews made by the customers. After a customer purchases the product from Olist Store, a seller gets notified to fulfil that order. Once the customer receives the product, or the estimated delivery date is due, the customer gets a satisfaction survey by email where he can give a note for the purchase experience and write down some comments. |
| Orders  (olist\_order\_dataset) | This is the core dataset. From each order, you might find all other information. |
| Products  (olist\_products\_dataset) | This dataset includes data about the products sold by Olist. |
| Sellers  (olist\_sellers\_dataset) | This dataset includes data about the sellers that fulfilled orders made at Olist. It is used to find the seller location and to identify which seller fulfilled each product. |
| Category Name (product\_category\_name\_transalation) | This dataset translates the product category name to English. |

*Table 4: Data Tables*



*Figure 4: Data Schema*

# **5. Methodology**



*Figure 5: Overview of project methodology*

The above figure, Figure 5, outlines the overarching process undertaken by the team to explore the best combination of segmentation modelling and classification modelling for our dataset.

The following subsections will explain, in greater detail, the actions taken at each step of the process outlined in Figure 5.

## 

## **5.1 Data Preprocessing**

### **5.1.1 Merging Data and Preliminary Table Selection**

| **Original Data Schema** | **Preliminary Table Selection** |
| --- | --- |
|  |  |

*Figure 6: Data Schema before and after Preliminary Table Selection*

Given our objective of segmenting consumers, the team has chosen to use only the following five tables:

1. Customers
2. Order Payments
3. Order Items
4. Order Reviews
5. Product Category Name Translation

Sellers dataset was deemed irrelevant to the segmentation of customers while the geolocation dataset which contained zip code, longitude and latitude data was deemed unnecessary given that the customer dataset already contains the customer’s state and city.

### **5.1.2 Exploratory Data Analysis and Preliminary Feature Selection**

|  | Top 10 states in Brazil based on the payment sum |
| --- | --- |
|  | Most customers make their purchase from **11am to 3pm**. |
|  | Number of orders made on this ecommerce platform has been steadily growing from the period of Sept 2016 to Sept 2018. However, we are unsure of the cause of the sudden drop in orders in the month of Sept 2018. |
|  | Most customers are **not frequent repeat customers**. Despite a few customers coming back to the store multiple times, their total spending is not very high. |
|  | There does not seem to be any obvious relationship between how recently the customer has visited the shop and their total amount spent |
|  | Most customers do opt for instalments for their payment. These instalments are mostly 10 or less. |

*Table 5: An overview the observation and insights based on EDA*

Continuing from above, we have conducted a preliminary feature selection of 21 variables. This also serves to reduce dimensionality, reducing the overall number of variables from 37 to 21.   
  
The following table briefly summarises the variables selected:

| **Data table** | **Attribute** | **Description** | **Example** |
| --- | --- | --- | --- |
| **Customers** | customer\_id | Key to the order dataset. | **customer\_id:**  9ef432eb6251297304e76186b10a928d |
| customer\_state | Customer’s state | **customer\_state:** SP |
| customer\_unique\_id | Unique identifier of a customer. | **customer\_unique\_id:** 861eff4711a542e4b93843c6dd7febb0 |
| **Order Payments** | order\_id | Unique identifier of the order. | **order\_id:** e481f51cbdc54678b7cc49136f2d6af7 |
| payment\_type | Method of payment chosen by the customer.  There are 4 payment types - credit card, voucher, boleto and debit card. A customer can choose more than 1 payment types for each purchase | **payment\_type=**credit card |
| payment\_installments | No. of instalments chosen by the customer | **payment\_installments:** 1 |
| payment\_value | The transaction value - the amount a customer needs to pay. | **payment\_value:** 38.71 |
| **Order Items** | order\_purchase\_timestamp | Customer’s purchase timestamp. | **order\_purchase\_timestamp:**  2017-10-02 10:56:33 |
| **Products** | product\_category\_name\_enlish | The category names of the products in english. | **product\_category\_name\_english :**health\_beauty |
| **Reviews** | review\_score | Note ranging from 1 to 5 (integer) given by the customer on a satisfaction survey. | **review\_score:** 5 |
| **Category Name** | product\_category\_name | The original category names of the products in portuguese. | **product\_category\_name:** beleza\_saude |
| product\_category\_name\_english | The category names of the products in english. | **product\_category\_name\_english:** health\_beauty |

*Table 6: Preliminary Feature Selection*

Apart from the unique identifiers, the above attributes were chosen because they were deemed most relevant to the objective of segmenting consumers. ‘payment\_type’ gives us information about consumers' preferred method of payment. More importantly, some payment methods are more profitable than others. For example, debit card payments provide better cash flow than credit purchases while credit cards can enable consumers to make higher value purchases. Similarly, ‘payment\_installments’ also acts as an indicator of purchasing power and allows us to analyse impact to cash flow. ‘payment\_value’ allows us to calculate the total amount each customer spends. Order purchase timestamp allows us to calculate recency and frequency of purchases which will be further elaborated in section 5.2.1. Product categories allow us to identify what types of products different consumer segments tend to purchase while Reviews give us insight to their satisfaction level. Lastly, the preliminary feature selection was also undertaken to reduce dimensionality and noise by irrelevant variables.

### **5.1.3 Data Cleaning**

Our dataset after merging and preliminary feature selection contained no null or missing values but had duplicated records that were easily removed. There were also no structural errors or any other data quality issues.

## **5.2 Data Transformation**

**5.2.1 Grouping and Categorical Data Encoding**

Each row of the dataset represents a single transaction. Given our goal of segmenting consumers at the customer level, we needed to transform the dataset to display data at a customer level. Furthermore, the literature review conducted showed most studies utilising customer level data. This would require an aggregation of the data. Numeric variables were aggregated by their means or sum as appropriate. For categorical variables such as payment type, one hot encoding was undertaken to allow for aggregation and accurate capturing of multiple payment types used by each customer.

Product categories, however, represented a challenge as the variable contained 71 unique values and utilising one-hot encoding would add 71 columns to the dataset, increasing dimensionality exponentially. To avoid the curse of dimensionality while still retaining the information provided by the product category column, for each customer, we only kept the product category of their most expensive purchase. This still provides us with insight on the most profitable product categories without increasing the number of dimensions.

### **5.2.**2 **Feature Engineering (RFM)**

Building on the works of Wu and colleagues in 2020 (Refer to Section 3.2) and further explained in Section 5.4.1, our team created 3 additional variables: Recency (R), Frequency (F) and Monetary (M).

Recency variable represents the number of days since the customer has made a purchase. It was calculated by subtracting the date of the most recent purchase made by the customer from the most recent purchase made by the entire dataset. Frequency represents the number of purchases a customer has made. This was derived by calculating the number of unique orders made by each customer. Monetary represents the total amount spent by each customer and is simply the sum of the payment value of each order.

### 

### **5.2.**3 **Normalisation**

For models that utilise continuous numeric variables, these continuous variables were first normalised by removing their mean and scaling to unit variance. This ensures that Variables that are measured at different scales contribute equally to the model fitting & model learned function and does not end up creating a bias for variables with larger scales.

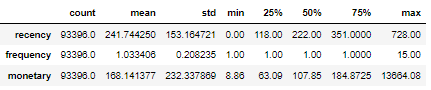
**5.1.4 Outlier Analysis and Removal**

Next, we looked at identifying and removing outliers. Outlier removal was done after Data Transformation to allow us to utilise the RFM variables to identify outliers.

| **Recency** | **Frequency** | **Monetary** |
| --- | --- | --- |
|  |  |  |

*Figure 7: Boxplot of RFM variables*

As seen in Figure 7, Frequency and Monetary variables are heavily right skewed with multiple observations beyond the upper whisker. We also identified customers that had RFM values that were greater than three standard deviations away from the mean.

  
*Figure 8: Statistical Summary of RFM variables*

However, looking at Figure 8, we can see that the minimum and maximum values for Recency and Frequency are actually reasonable. A recency value of 0 and 728 indicates that a customer purchased 0 days ago and another bought 728 days ago respectively. A frequency value of 1 and 15 indicates that a customer bought once and another bought 15 times. These values are reasonable and should not be considered as outliers or excluded from the segmentation modelling.

The minimum and maximum value of the monetary variable, R$8.86 and R$13,664, translates to SGD2.50 and SGD3790 respectively. This too may be reasonable but further investigation is necessary. The most expensive item bought on the e-commerce platform, Shopee, was a foldable solid table worth SGD2,500 (Shopee, 2017). As such, we looked at purchases that were above SGD2,500 (R$8,992).   
  
There was only a single transaction, which was the max value for the monetary column and the purchase was for a fixed landline telephone. R$13,664 for a fixed landline telephone is extreme and much higher than the mean landline price of R$309. Hence, this observation was removed. The new maximum value for the monetary column was R$708 which was well within the reasonable range. Therefore, no other outliers were identified or removed.

## **5.3 Data Reduction**

### **5.3.1** Principal Component Analysis **(PCA)**

Similar to normalisation, for models that utilise continuous variables, Principal Component Analysis (PCA) was undertaken to reduce dimensionality. Principal component analysis (PCA) is a popular technique for analysing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while preserving the maximum amount of information, and enabling the visualisation of multidimensional data. The way PCA is applied is further elaborated in the subsequent sections for the specific models.

## 

## 5.4 Customer Segmentation

Upon researching on the different methods to conduct segmentation, this is the summarised methods in the table below:

| **Segmentation Method** | **Main factors considered in segmentation** | **Advantages** | **Disadvantages** |
| --- | --- | --- | --- |
| **A priori segmentation** | Widely available demographic characteristics (age, gender, location, employment, etc.) | Simple, quick, uses easily accessible data | Can be too simplistic to confer real-world value |
| **Needs-based segmentation** | Customers’ needs or drivers | Relevance: Better grouping based on real needs in relation to your products/services | Requires market research (higher investment costs) |
| **Cluster-based segmentation** | Multiple possible factors, as identified through mathematical analyses | Reduces bias via the use of objective data, expands segmentation possibilities significantly | Requires market research and further statistical analysis, most likely requiring third-party involvement (higher investment costs) |
| **RFM segmentation** | Recency and frequency of customer interactions, economic/monetary value | Provides fairly strong insights, based on easy and understandable yet readily accessible data, easy to implement and interpret | Still may miss important characteristics that could help with further segmentation |

*Table 7:The different methods to conduct segmentation (Ada, 2022)*

Our team will be exploring the following three methods of segmentation:

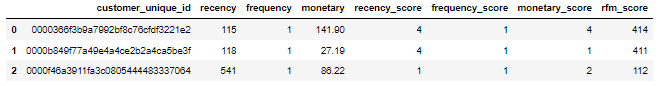
1. RFM scoring
2. K-Means clustering using continuous variables (Cluster-based)
3. K-Prototype using all our variables - both continuous and categorical variables including the RFM values. (Cluster-based)

This is largely because we are hoping to build on the works of Wu et al. (2020) (Refer to Section 3.2) who combined K-Means with RFM but failed to consider RFM individually and did not utilise additional continuous variables or categorical variables apart from RFM. Additionally, the above methods were best suited for our dataset that did not include much demographic, behavioural or psychographic data of consumers.

### **5.4.1 RFM Scoring**

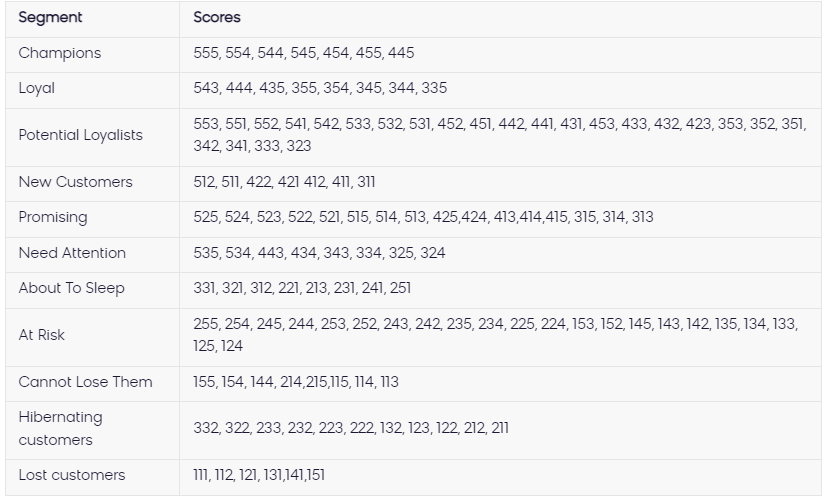
RFM segmentation is an existing business segmentation method which categorises customers into different segments based on their scores on 3 main attributes - Recency, Frequency, and Monetary. The RFM methodology was built on the hypothesis that customers that are most loyal and valuable would be those that ordered most recently, most frequently and spent the highest amount of money (Bult & Wansbeek, 1995). Recency is derived from the number of days since last purchase, Frequency is the number of times a specific customer has made a purchase, and monetary is the total sum of money the customer has spent. While more than 50 methods exist on how to obtain the RFM scores (Wei et al.,2010), we undertook one of the most popular and straightforward methods, scoring customers by discretizing their RFM values into 5 quartiles/quintiles (segments) of equal size.

Thus, each of these RFM parameters are assigned a score in the range of 1-5 with 5 being the highest score. For example, if a customer has made a purchase recently, and has a record of making many expensive purchases, they would be given a RFM score of 555 (5 for recency, 5 for frequency, and 5 for monetary).



*Figure 9: Sample of RFM scoring taken from code*

With the rfm\_score that we have given for each customer, we were able to map them to 11 different customer segments based on the model developed by Bloomreach.



*Figure 10: Customer Segments based on RFM Score (Bloomreach, 2022)*

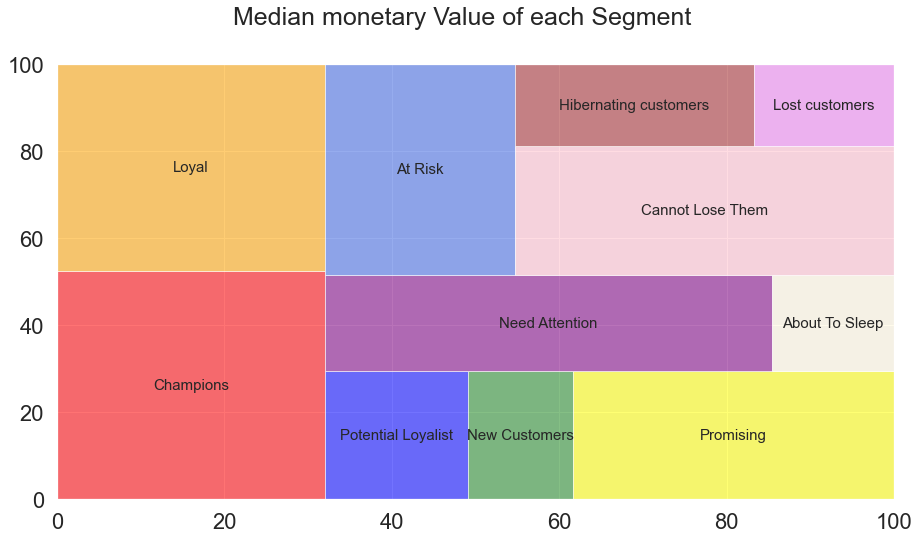
According to these segments, the best customer segments for business would be the “Champions” category, where these customers spent the most amount of money, have bought most recently, and have also bought most recently. There are also descriptive characteristics as well as actionable steps for these customer profiles as shown below.



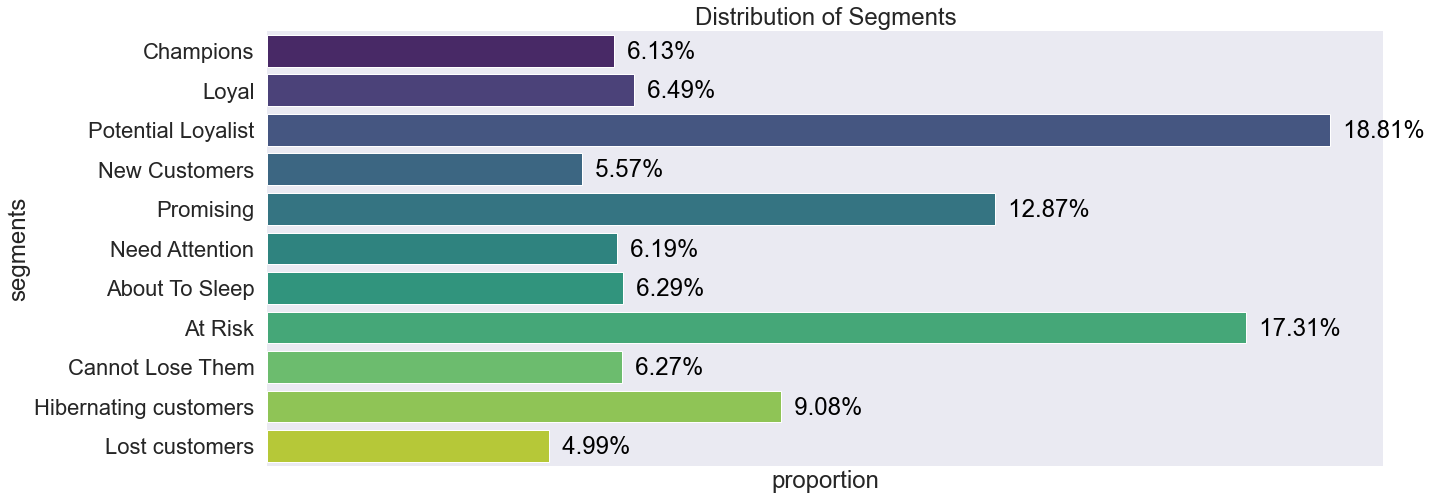
*Figure 11: Customer Segment description (Bloomreach, 2022)*

Based on Figure 11, businesses can have further insight on their customer groups and channel more of their resources on addressing specific customer segments.

Applying these method to our dataset, we have obtained the following results:



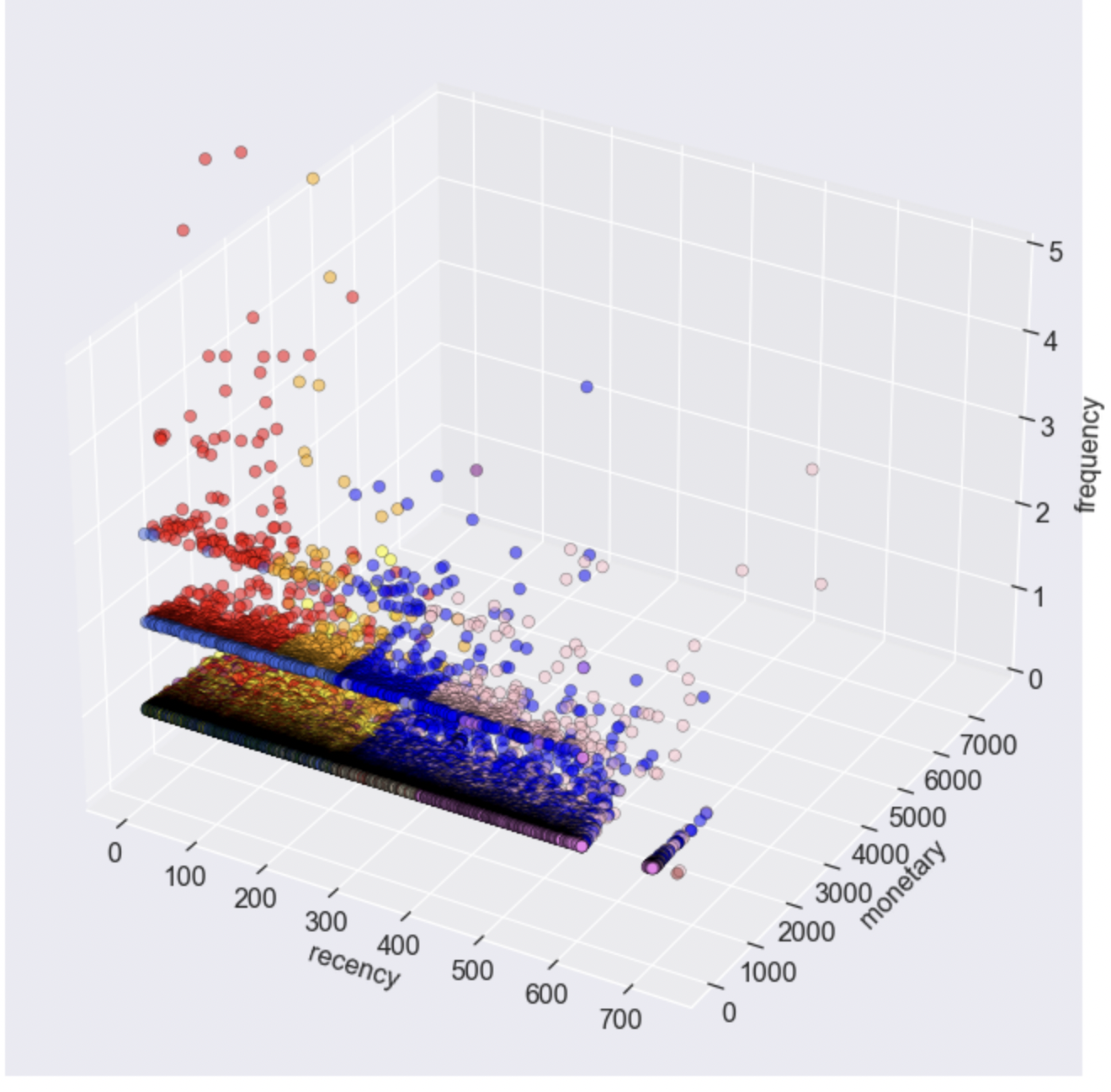
*Figure 12: Customer Segment Value Treemap*

**

*Figure 13: Customer Segment Distribution*

Figure 12 illustrates the median monetary value that each segment brings to the business through the size of each rectangle, while Figure 13 shows the distribution of the 11 consumer segments in our data set. From Figure 12, we can see how “Champions'' are the most valuable customer segment. Figure 13, shows that Olist has a high proportion of “Potential Loyalist '' and “At Risk” customers. These figures suggest a volatile customer base for the business, which can indicate that more efforts need to be made to improve the customer acquisition process and seize the opportunity of converting potentially loyal customers into loyal ones, and better retention efforts so that customers at risk would not be lost. This clearly shows the ease of use and strong actionable insights that the RFM method provides.

Upon plotting our resulting customer segments in a 3d graph, we have obtained the following results as shown below.



*Figure 14: 3D plot of customer segments*

For computational efficiency, our group used the Davies-Bouldin Index & the Calinski-Harabasz Index for our cluster evaluation. These scoring methods are available and were obtained from the sklearn.metrics library.

The Calinski Harabasz Score is essentially the ratio of the sum of between-cluster dispersion and within-cluster dispersion (Calinsky & Harabasz, 1974). This means that a higher value of the calinski score indicates better clustering since it represents a higher value of between-cluster dispersion and a lower value of within-cluster dispersion which we are seeking.

The Davies-Bouldin Index Score is defined as the average similarity measure of each cluster with its most similar cluster (Davies & Bouldin, 1979). The similarity is the ratio of within-cluster distances to between-cluster distances. Thus, lower values indicate that clusters are farther apart and less dispersed which reflects better clustering.

**Results and Discussion**

The scores for our RFM clustering are as follows:

**Calinski Harabasz Score: 4590.05**

**Davies-Bouldin Index Score: 4.35**

Despite its ease of use and strong actionable insights, the RFM method has some major drawbacks.. One major drawback of the RFM method is that it forces customers to fit into a set number of segments. In this case, it is 11 segments. Given the dynamic nature of real world data, 11 segments may not be the optimal number of customer segments given a specific dataset. More importantly, only considering 3 features (RFM) is incredibly limiting as other meaningful data would be missed such as geographical location, product categories and other demographic or psychographic factors which could influence the future purchase behaviour of customers.

Keeping these limitations in mind, we continue to explore the use of clustering algorithms that incorporate more variables and provide a better notion of vector space quantization. We continue to use this method as a feature engineering technique, and decided to include the RFM features alongside other variables to train the clustering models

### **5.4.2 K Means Clustering based on continuous variables**

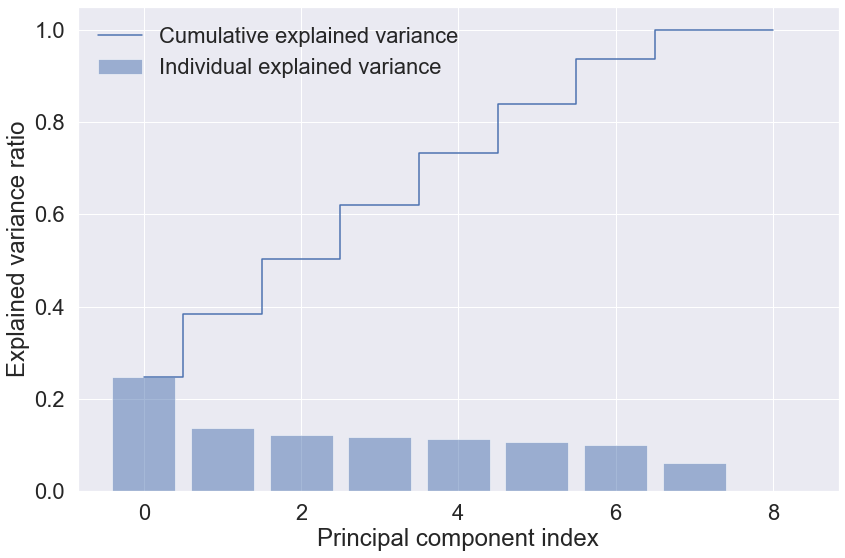
Since K Means uses distance-based measurements to determine the similarity between data points, we only select continuous variables as inputs for this clustering method.

Continuous variables:

* Order Recency (R)
* Frequency of Orders (F)
* Total Monetary Value (M)
* Mean No. Payment Instalments Required (Lower = Better Cash Flow for company)
* Review Score (Proxy for customer satisfaction)
* Proportion of Purchases paid with Credit Card
* Proportion of Purchases paid with Vouchers
* Proportion of Purchases paid with Debit Card
* Proportion of Purchases paid withBoleto

To ensure all variables are given equal weight and that the algorithm does not bias to any specific variable because of its scale, we first standardise the data by subtracting the mean and scaling to unit variance.

We also experimented building the model with and without the use of PCA to reduce dimensions.For the PCA approach, we performed PCA on the 9 continuous variables mentioned above.



*Figure 15: K-Means PCA Explained Variance Ratio Plot*

6 principle components (PC) were selected that explain 83.9% of the total variation. While there is no strict rule or guidelines in terms of selecting the number of PCs, we internally decided on a threshold of 80% or more of the total variation. This allows us to retain a significant amount of variance in the data while still reducing the number of dimensions.

The number of clusters (k) is the most important hyperparameter in K-Means clustering. We find k using the elbow method built using a Distortion Score (Sum of Squared Errors) and the Calinski Harabasz Score. The following tables show the optimal number of clusters, the Sum of Squared Errors (SSE) and the Calinski Harabasz Score for the model with PCA and without PCA:

**K-Means clustering: With PCA**

|  | |
| --- | --- |
| **Optimal No. of Clusters** | 7 |
| **SSE** | 168,482.56 |
| **Calinski Harabasz Score** | 43,176.24 |

**K-Means clustering: Without PCA**

|  | |
| --- | --- |
| **Optimal No. of Clusters** | 6 |
| **SSE** | 289,091.29 |
| **Calinski Harabasz Score** | 29,676.46 |

**Results and Discussion**

**K-Means Clustering: With PCA**

**SSE: 168,482.56**

**Calinski Harabasz Score: 43,176.24**

**Davies Bouldin Score: 0.78**

The K-Means Model **with PCA** was selected as it has a lower SSE and a higher Calinski Harabasz Score. The K-Means Model **with PCA** also outperforms the RFM segmentation with a higher Calinski Harabasz Score and a lower Davies Bouldin Score.

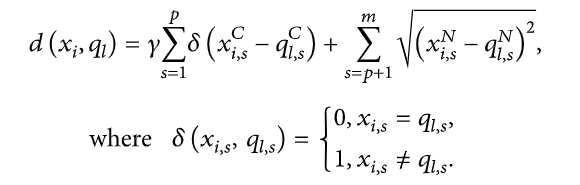
Nevertheless, the team notes the limitations of K-Means clustering which includes the inability to consider categorical variables. We continue to explore K-Prototypes, a clustering algorithm that is able to incorporate both continuous numerical and categorical data.

### **5.4.3 K**-**Prototype Clustering based on both categorical and numerical variables**

In order to have a full understanding of the data including both categorical and numerical attributes, we have decided to use k-prototype (k-prototype = k-means + k-modes). Huang (Nguyen, 2017) developed the k-modes algorithm which is an extension from k-means algorithm by using:

1. A simple matching dissimilarity measure for categorical attributes
2. Modes in place of means for clustering
3. Frequency-related strategy to update modes to minimise the clustering cost

This algorithm is a hybrid clustering algorithm based on partitioning and takes into account the Dissimilarity Coefficient of Categorical Feature and Numerical Feature at the same time, as shown in the following formula:

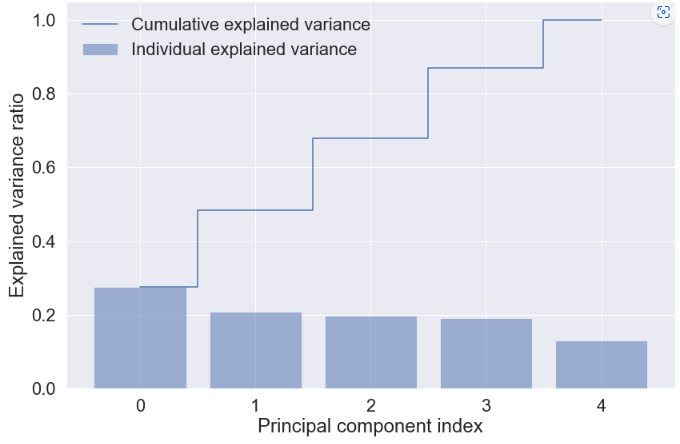


*Figure 16: K-Prototype formula (Jia & Song, 2020)*

Selecting ‘Huang’ as the init, the model will select the first k distinct objects from the data set as initial k-modes and then assign the most frequent categories equally to the initial k-modes. The ‘Cao’ approach selects prototypes for each data object based on the density of the data point and the dissimilarity value. Studies have shown that ‘Cao’ as it is considered to be a better method to initialise cluster centres as compared to Huang which we have adopted in our analysis (Nguyen, 2017). Our dataset consists more of categorical values which have been one-hot encoded and are more dense compared to a dataset containing only continuous attributes. With this in consideration, our group has decided to use ‘Cao’ as the init value.

In terms of variable selection for K Prototype clustering, the continuous variables selected are recency, frequency, monetary, review score and mean number of instalments required. Categorical Variables are payment methods such as Credit Card, Voucher, Boleto & Debit Card (1: Used, 0: Never Used), Customer state and Most Expensive Product Category purchased.

We then normalised the continuous variables and performed PCA, given that the K-Means model with PCA outperformed the model without PCA. After which, we selected 4 principle components that explained 87% of variation. Similar to the approach taken for the K-Means model, we set an internal threshold of 80% which allows us to retain a significant amount of variance in the data while still reducing the number of dimensions.

  
*Figure 17: K-Prototype PCA Explained Variance Ratio*

One major drawback of the K-Prototype algorithm is its extremely long runtime. Running the model once on our data set with approximately 93k rows and 10 columns took 41 minutes. As such, due to time and resource constraints, we bypassed the elbow method to select the optimal number of clusters. Instead, we selected the optimal number of clusters based on the previous best k-means model where k = 7.

**K-Prototype**

| **3-D Visualisation of Clusters (PC 1, PC 2,PC 4)** | **2-D Visualisation of Clusters (PC 1 & PC 2)** |
| --- | --- |
| **Distribution of Segments/Clusters** | |
| **Calinski Harabasz Score** | 33,828.89 |
| **Davies-Bouldin Index Score** | 1.68 |

### **5.4.4 Cluster/Segmentation Evaluation**

Comparing the 3 clustering methods selected of RFM, K-Means and K-Prototype, the models are evaluated based on Calinski Harabasz score and Davies-Bouldin Index score. As K-Means has the highest Calinski Harabasz score and the lowest Davies-Bouldin Index score, we identified the K-Means model as the best and final clustering method to segment and label our dataset.

|  | **RFM Scoring** | **K-Means** | **K-Prototypes** |
| --- | --- | --- | --- |
| **Calinski Harabasz Score** | 4590.05 | 43,176.24 | 33,828.89 |
| **Davies-Bouldin Index Score** | 4.34 | 0.78 | 1.68 |
| **Pros** | * Easy and efficient to compute * Only requires 3 variables * Easy to interpret * Highly actionable | * Better notion of vector space quantization * Incorporates RFM variables along with other continuous variables | * Better notion of vector space quantization * Able to utilize both numerical and categorical data |
| **Cons** | * Force fits consumers into 11 fixed segments * Doesn’t take into consideration other variables   + Eg. Satisfaction, Geolocation, other purchase preferences | * Doesn’t take into account categorical variables * Clusters more difficult to interpret * Less actionable | * Extremely long runtime * Clusters more difficult to interpret * Less actionable |

*Table 8 : Comparison of RFM, K-Means and K-Prototypes*

With that being said, each method has its own pros and cons. Starting with RFM scoring, it was easy and efficient to compute, only requiring 3 variables, easy to interpret and is also highly actionable. However, it forcefully fits consumers into the 11 fixed segments. It also does not take into account other variables such as satisfaction, geolocation and other purchase preferences.

Moving on to K-means, its advantages includes, better notion of vector space quantization and incorporates RFM variables along with other continuous variables. However, it does not take into account categorical variables, clusters are more difficult to interpret and it is less actionable.

Lastly, similar to K-Means, K-prototype also provides a better notion of vector space quantization as compared to RFM. K-Prototype also results in clusters that are not easy to interpret and are less actionable. More importantly, unlike K-Means and RFM, K-Prototype allows for the use of both numerical and categorical data which could be significantly useful in certain circumstances. However, its long runtime is a major drawback and one that would make it less user-friendly for firms with limited computing capabilities.

### 

### **5.4.5 Final Clusters after chosen K-Means clustering method**

|  | **customer\_state** | | **recency** | **frequency** | **monetary** | **mean\_review\_score** | **boleto** | **credit\_card** | **debit\_card** | **voucher** | **most\_exp\_category** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | count | mode | median | median | median | mean | mean | mean | mean | mean | mode |
| **0** | 21681 | SP | 395.0 | 1.0 | 125.520 | 4.62 | 0.00 | 0.99 | 0.00 | 0.00 | bed\_bath\_table |
| **1** | 33658 | SP | 137.0 | 1.0 | 92.780 | 4.63 | 0.00 | 0.99 | 0.00 | 0.00 | health\_beauty |
| **2** | 18190 | SP | 235.0 | 1.0 | 93.83 | 4.11 | 0.99 | 0.00 | 0.00 | 0.00 | health\_beauty |
| **3** | 3462 | SP | 246.5 | 1.0 | 86.73 | 4.08 | 0.00 | 0.28 | 0.00 | 0.71 | bed\_bath\_table |
| **4** | 1422 | SP | 130.0 | 1.0 | 89.69 | 4.18 | 0.00 | 0.00 | 1.00 | 0.00 | health\_beauty |
| **5** | 3937 | SP | 186.0 | 2.0 | 464.07 | 4.18 | 0.10 | 0.88 | 0.01 | 0.01 | bed\_bath\_table |
| **6** | 11045 | SP | 210.0 | 1.0 | 133.38 | 1.48 | 0.00 | 0.99 | 0.00 | 0.00 | bed\_bath\_table |

*Table 9 : Statistical Summary of Final Clusters*

| Cluster | **Customer Profile** |
| --- | --- |
| 0 | **Lost Customers:** Haven’t purchased in a year |
| 1 | **Core:** Majority of Customer Base**,** Highly satisfied, Average Spending, Pay mostly using Credit Card |
| 2 | **Boleto Users:** Purchase mainly using Boleto, Average spending, Purchased relatively long ago |
| 3 | **Bargain Hunters:** Purchase using vouchers |
| 4 | **Promising**: Satisfied customers that purchased not too long ago |
| 5 | **Champions: (Most Valuable Segment )** Spends the highest, most frequently. Satisfied and mainly purchases with credit card |
| 6 | **Dissatisfied Customers:** Poor Review Scores |

*Table 10 : Interpretation of Clusters*

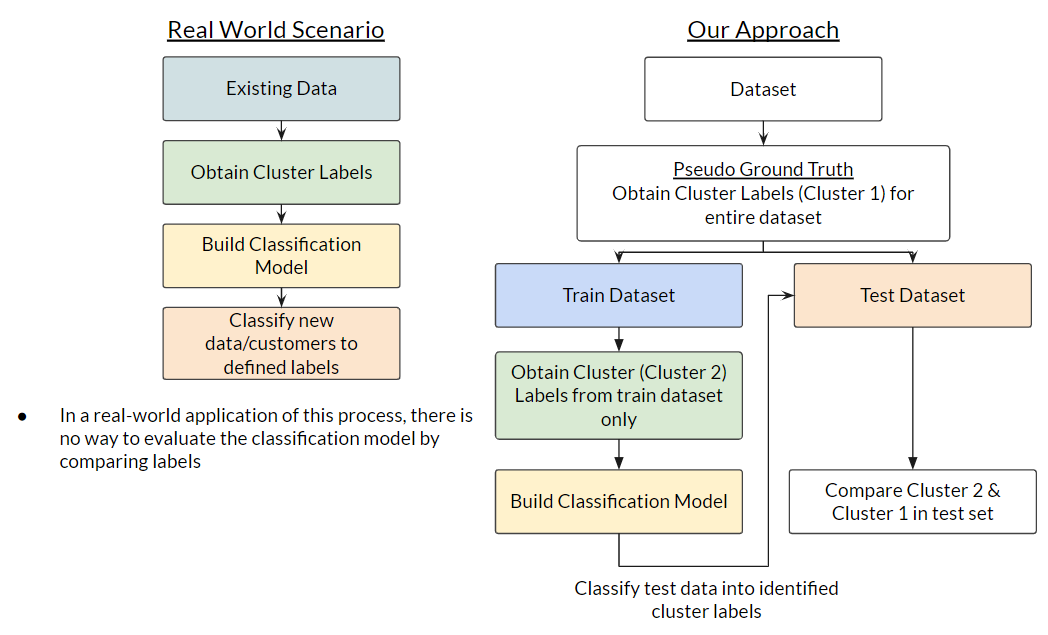
From the 2 tables above, we can interpret the final clusters produced by the K-Means model. Given that the purpose of this project is not to actually analyse the consumer segments of Olist customers, we would not be analysing the customer profiles any further. The tables above are displayed simply to show that it is possible to profile customer segments created via clustering algorithms using business domain knowledge. How it is done, however, is up to the firms themselves.

## 5.5 Classification Modelling

Next, we build a classification model on the labels obtained from clustering. In a real world scenario, there is no way to measure the accuracy of the classification model, given that the classification model will be used to predict on new customer data that do not have ground truth labels.

As such, we have adopted another approach by clustering the entire dataset to obtain cluster\_1 which will act as the Pseudo Ground Truth. We then divided the data into train and test sets. Using the train dataset, we then ran the algorithm again to obtain another set of labels, cluster\_2. We trained a classification model on the train dataset and utilised cluster\_2 as the target variable. This simulates building a classification model on existing data in a real world scenario.

After which, we will use this model to predict the cluster labels for the test data which simulates predicting the segments of new customers (unseen data) in the real world scenario. However, with this approach, we are able to compare the cluster labels (cluster\_2 vs cluster\_1) and evaluate the classification model. This is summarised in Figure 18 below:



*Figure 18 : Classification Methodology*

Building on our literature review, we explored three classification algorithms: Decision Trees, Random Forest and XGBoost.

### 5.5.1 Decision Tree

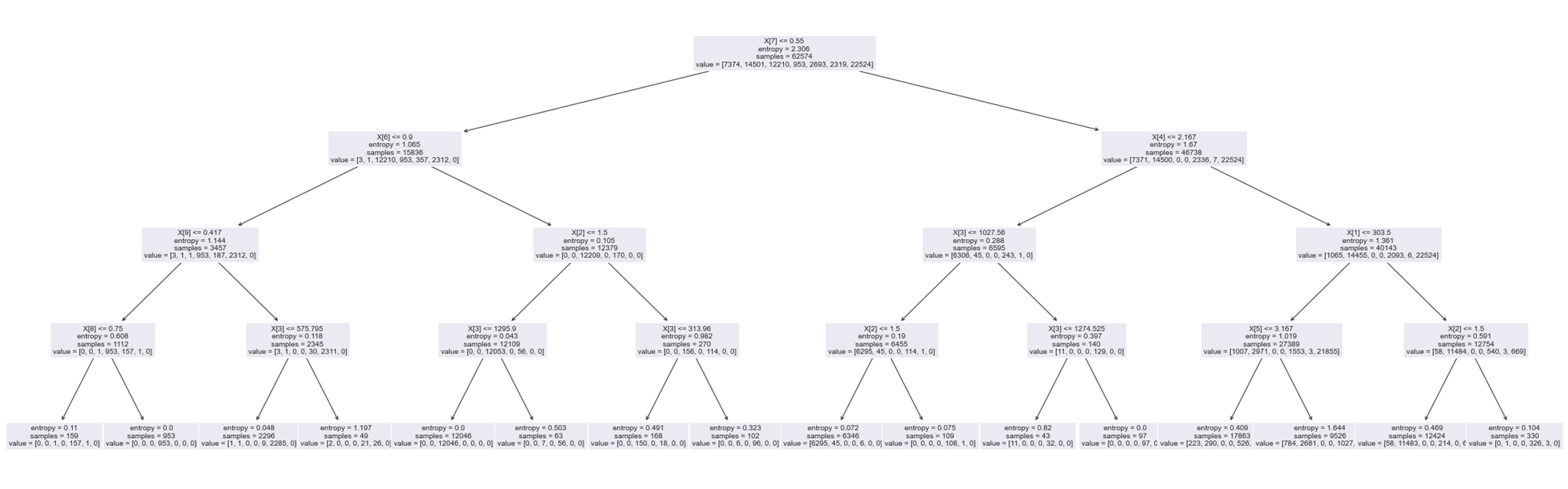
**Feature X:**

* Recency
* Frequency
* Monetary
* Mean\_num\_installments
* Mean\_review\_score
* Boleto
* Credit\_card
* Debit\_card
* Voucher

**Target Y:**

* Cluster\_2

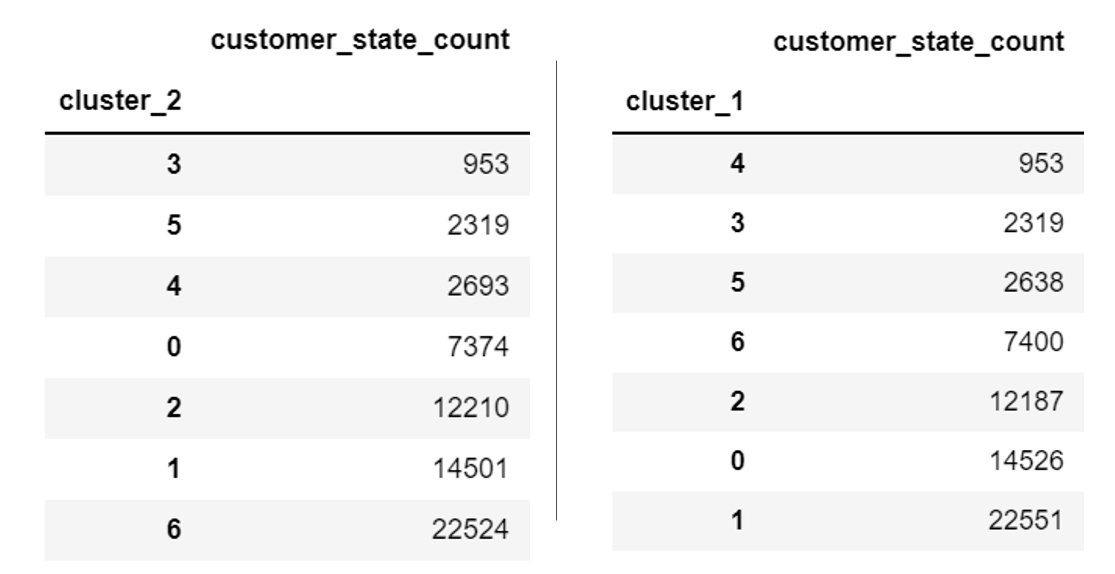
Using the training dataset, maximum depth of 4 and random state of 42, we initialised and trained the Decision Tree Classifier.



*Figure 19: Decision Tree*

After which, we predicted the clusters of test data, labelled as cluster\_2. Since we are taking cluster columns as our pseudo cluster truth, we compared cluster\_2 with cluster\_1 to remap the cluster labels. The purpose of this is because the cluster labels may not always be the same each time we run the clustering model since we are unable to instruct the K-means implementation to use the same names for the clusters each time. In our case, we are using cluster size ranking as a proxy.

For example, comparing the two clusters’ label as shown in Figure 20, we have mapped the clusters as cluster\_map = {3:4, 5:3, 4:5, 0:6, 2:2, 1:0, 6:1} whereby cluster 3 in cluster\_2 column is mapped to cluster 4 in cluster\_1 column.



*Figure 20: Comparing the cluster labels between cluster\_2 and cluster\_1 column*

accuracy: 0.892

precision: 0.892

recall: 0.892

f1 score: 0.892

### **5.5.2 Random Forest**

Using the same training and testing datasets as in the Decision Tree, we first initialise the Random Forest Classifier with n\_estimators=100 and random state of 42. Upon training and running prediction, the following are the metrics for our Random Forest Classifier.

accuracy: 0.991

precision: 0.991

recall: 0.991

f1 score: 0.991

### **5.5.3 XGBoost**

Similarly to running both the Decision Tree and Random Forest classification, we ran the XGBoost model. It is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework, producing the following evaluation metrics:

accuracy: 0.991

precision: 0.991

recall: 0.991

f1 score: 0.991

### 5.5.4 Classification Model Results & Discussion

Results from the three classification models:

|  | **Decision Tree** | **Random Forest** | **XGBoost** |
| --- | --- | --- | --- |
| **Accuracy** | 0.89179 | 0.99114 | 0.99114 |
| **Precision** | 0.89179 | 0.99114 | 0.99114 |
| **Recall** | 0.89179 | 0.99114 | 0.99114 |
| **F1 Score** | 0.89179 | 0.99114 | 0.99114 |

*Table 11 : Comparison of Decision Tree, Random Forest and XGBoost*

Evaluating the results above, we noticed that both Random Forest and XGBoost produced a better result as compared to the Decision Tree classification model.

However, since both classification models give us similar results we had to consider the difference between the two. Based on our dataset, we would prefer Random Forest over XGBoost for three reasons. Firstly, Random Forest gives us more preferences for hyperparameters, making it easier to tune and optimise the model. Secondly, it is less likely to overfit the data because we are creating multiple trees, with each tree being trained differently. The decisions of each tree are then combined to make the final classification. Lastly, Random Forest has many trees with leaves of equal weight so that high accuracy and precision can be obtained easily with the available data which XGBoost does not have.

# **7. Conclusion & Future Work**

## **7.1 Conclusion**

In conclusion, we ran the three methods for clustering - RFM scoring, k-means and k-prototype to determine the best segmentation method before running the classification models. From our working model, K-Means and K-Prototype outperformed RFM scoring to form customer segments. Our project further strengthens the argument that marketers should combine modern data mining techniques with their domain expertise to determine the best segment consumers, rather than relying on business models alone. While our work has shown that K-Means performed the best, ultimately, the best model is highly dependent on its specific business use case, data quality, as proven by Coussement et al. (2014), and data availability.

Our project has also shown the possibility of leveraging both clustering and classification algorithms to segment consumers more responsively. By building a classification model on a defined cluster labels or segments, firms can classify new consumers or reclassify existing consumers into their value segments without having to run the entire clustering algorithm again. This also allows the firms to predict the customer value in real-time, continuously, so they can employ highly targeted and personalised marketing tactics to each individual customer. This would likely lead to greater return on marketing spend, more satisfied customers and potentially stronger customer acquisition and retention.

Additionally, for companies that have useful categorical variables, our work has shown that the K-Prototype algorithm is well worth exploring, especially for firms that have the computational means. On top of that, firms can also explore a more computationally efficient strategy of building a K-Prototype model on a subset of the data and then leveraging a classification model to label the rest of the data.

## **7.2 Future Work**

We have also identified future work that can be explored by experts from different domains.

### 7.2.1 Businesses / Marketers

Firstly, businesses or marketers can explore the more efficient and dynamic process of customer segmentation that was proposed in this paper in the real world and measure the actual potential return on investment, similar to how Wu et al. (2020) implemented their marketing strategies in the real world. This may prove to be a better way to evaluate the process rather than simply relying on evaluation metrics.

Secondly, we propose that organisations should focus less on utilising the most advanced algorithms available or best algorithms shown in literature but rather focus on identifying the best algorithm for their use case based on data quality, accessibility and availability.

### 7.2.2 Academia / Research

For researchers, it appears algorithms to cluster data using both categorical and continuous variables are still being explored. Future work should look into improving the computational efficiency of these algorithms or seek to compare and evaluate the different algorithms.

Our project also did not consider density based clustering algorithms such as DBSCAN or HDBSCAN. Further research can explore how these algorithms perform for consumer segmentation and how they interact with RFM variables.

# **8. References**

Ada, T. (2022, April 21). Customer Segmentation: Why is it Important to Segment Your Customers? Ada Support. Retrieved November 5, 2022, from <https://www.ada.cx/customer-segmentation>

Bain, Google, & Temasek. (2022, November 8). *E-conomy sea 2022*. e-Conomy SEA 2022. Retrieved November 18, 2022, from https://www.bain.com/insights/e-conomy-sea-2022/

Bhandari, A. (2020, June 16). *AUC-ROC Curve in Machine Learning Clearly Explained*. Analytics Vidhya. Retrieved November 5, 2022, from <https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/>

Bloomreach. (n.d.). *RFM segmentation*. Engagement. Retrieved November 18, 2022, from https://documentation.bloomreach.com/engagement/docs/rfm-segmentation#rfm-segments

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science*, *14*(4), 378–394. https://doi.org/10.1287/mksc.14.4.378

Chaudhuri, N., Gupta, G., Vamsi, V., & Bose, I. (2021, June 7). *On the platform but will they buy? Predicting customers' purchase behaviour using deep learning*. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0167923621001329

Coussement, K., Van den Bossche, F. A.M., & De Bock, K. W. (2012, October 5). *Data accuracy's impact on segmentation performance: Benchmarking RFM Analysis, logistic regression, and decision trees*. ScienceDirect. Retrieved October 29, 2022, from <https://doi.org/10.1016/j.jbusres.2012.09.024>

Davies, David L.; Bouldin, Donald W. (1979). “A Cluster Separation Measure”. IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224-227

Dick, A.S., Basu, K. Customer loyalty: Toward an integrated conceptual framework. JAMS 22, 99–113 (1994). https://doi.org/10.1177/0092070394222001

Jia, Z., & Song, L. (2020, July 25). *Weighted k-Prototypes Clustering Algorithm Based on the Hybrid Dissimilarity Coefficient*. Hindawi. Retrieved November 5, 2022, from <https://www.hindawi.com/journals/mpe/2020/5143797/>

Jones, T. O., & Sasser, W. E. (2014, August 1). Why satisfied customers defect. Harvard Business Review. Retrieved November 18, 2022, from <https://hbr.org/1995/11/why-satisfied-customers-defect>

Martinez, A., Schmuck, C., Pereverzyez Jr., S., Pirker, C., & Haltmeier, M. (2022, January 17). *A machine learning framework for customer purchase prediction in the non-contractual setting*. ScienceDirect. Retrieved October 29, 2022, from https://doi.org/10.1016/j.ejor.2018.04.034

Nguyen, H. (2017, December 21). *Clustering Categorical Data Using Community Detection Techniques*. NCBI. Retrieved November 5, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753022/>

Shopee (2017, November 16). *In Singapore, the most EXPENSIVE item bought during this sale was a Foldable Solid Wood Table for $2,500!* Retrieved November 5, 2022, from <https://www.facebook.com/ShopeeSingapore/photos/a.967240393414233/967240793414193/?type=3>

T. Caliński & J Harabasz (1974) A dendrite method for cluster analysis, Communications in Statistics, 3:1, 1-27, DOI: 10.1080/03610927408827101

Wei, Jo-Ting & Lin, Shih-Yen & Wu, Hsin-Hung. (2010). A review of the application of RFM model. African Journal of Business Management December Special Review. 4. 4199-4206.

Wu, Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. Mathematical Problems in Engineering, 2020, 1–7. <https://doi.org/10.1155/2020/8884227>

Zazueta, Z. (2020, October 23). *K-Prototypes clustering — for when you're clustering continuous and categorical data*. Zach Zazueta. Retrieved November 5, 2022, from <https://zachary-a-zazueta.medium.com/k-prototypes-clustering-for-when-youre-clustering-continuous-and-categorical-data-6ea42c2ab2b9>