Text

Description automatically generated

BC2407 ANALYTICS II:  
ADVANCED PREDICTIVE ANALYTICS

Applications of Analytics to Business: The Jack (MA) of all Trade(RS)

AY21/22 SEM 2 | SEMINAR 1, TEAM 7  
Instructor: Prof Neumann Chew

|  |  |
| --- | --- |
| Matric No. | Group Member Name |
| U2010149F | Alvin Lim Swee Hung |
| U2021449D | Deionna Chee Rui Ping |
| U2021546F | Hor Chia Yee, Lara |
| U2021729K | Leow Ken Hing Bryan |
| U2040650E | Timmothy Yonathan |

**Table of Contents**

[EXECUTIVE SUMMARY 5](#_Toc99921505)

[Problem Statement 5](#_Toc99921506)

[Our Solution and Results 5](#_Toc99921507)

[Findings around the Business Problem 5](#_Toc99921508)

[Findings around our Analytics Models 5](#_Toc99921509)

[1. The Business Problem / Opportunity Statement 1](#_Toc99921510)

[1.1. The Opportunity Statement 1](#_Toc99921511)

[1.2. Justification of Business Problem 1](#_Toc99921512)

[1.2.1 Rise of E-Commerce 1](#_Toc99921513)

[1.2.2 Emergence of DTC Model 2](#_Toc99921514)

[1.2.2.1 The Appeal of DTC - Expanding and Enhancing Seller’s Presence 2](#_Toc99921515)

[1.2.2.2 The Appeal of DTC - Greater Opportunities to Build Positive Customer Relationships 2](#_Toc99921516)

[1.2.2.3 The Appeal of DTC - Increased Margins 2](#_Toc99921517)

[1.2.2.4 The increasing feasibility of managing a DTC model 2](#_Toc99921518)

[1.2.2.5 Case studies of brands that successfully adopted the DTC model, and DNVBs 2](#_Toc99921519)

[1.3 Desired Business Outcomes 3](#_Toc99921520)

[1.3.1 Cost-Effective Means to Expand and Enhance Seller Presence 3](#_Toc99921521)

[1.3.2 Fostering Positive Seller Impressions and Customer Relations 3](#_Toc99921522)

[1.3.3 Business Strategy Excellence through Understanding of Customer Lifetime Value (CLV) Trends and Drivers 3](#_Toc99921523)

[2. Data Selection and Cleaning 3](#_Toc99921524)

[2.1 Data Cleaning: UCI eCommerce 3](#_Toc99921525)

[2.1.1 Handling Missing Values 3](#_Toc99921526)

[2.1.2 Filtering Non-Purchase Transactions via InvoiceNo and Quantity 3](#_Toc99921527)

[2.1.3 Handling Cancelled Transactions 4](#_Toc99921528)

[2.1.4 Deriving a new ‘Product Variations’ column from StockCode 4](#_Toc99921529)

[2.2 Data Cleaning: Olist eCommerce 5](#_Toc99921530)

[2.2.1 Selecting “delivered” Status 5](#_Toc99921531)

[2.2.2 Merging Data 5](#_Toc99921532)

[2.2.3 Removing Duplicates 5](#_Toc99921533)

[2.2.4 Converting Time Data to the Right Format 5](#_Toc99921534)

[3. Expanding and Enhancing Seller Presence 5](#_Toc99921535)

[3.1. Introduction to the Analytics Solution: Association Rules 5](#_Toc99921536)

[3.2 Measures for Association Rules 6](#_Toc99921537)

[3.3. Association Rules on UCI Dataset 6](#_Toc99921538)

[3.3.1 Data Preparation 6](#_Toc99921539)

[3.3.2 Results 6](#_Toc99921540)

[3.3.3 Insights 6](#_Toc99921541)

[3.4. Association Rules on Olist Dataset 6](#_Toc99921542)

[3.4.1 Data Preparation 6](#_Toc99921543)

[3.4.2 Results 6](#_Toc99921544)

[3.4.3 Insights 7](#_Toc99921545)

[3.5. Expanding and Enhancing Seller Presence Through Association Rules 7](#_Toc99921546)

[4. Building Positive Customer Relationships 7](#_Toc99921547)

[4.1. Introduction to the Analytics Solutions 7](#_Toc99921548)

[4.1.1 Linear Regression 7](#_Toc99921549)

[4.1.2 Quantile Regression 8](#_Toc99921550)

[4.1.3 Multivariate Adaptive Regression Splines (MARS) 8](#_Toc99921551)

[4.2. Measures and Targets for Review Score Predictions 8](#_Toc99921552)

[4.3. Data Preparation 8](#_Toc99921553)

[4.4. Linear Regression 8](#_Toc99921554)

[4.4.1 Implementation 8](#_Toc99921555)

[4.4.2 Results 9](#_Toc99921556)

[4.4.3 Conclusion 9](#_Toc99921557)

[4.5. Quantile Regression 9](#_Toc99921558)

[4.5.1 Implementation 9](#_Toc99921559)

[4.5.2 Results 9](#_Toc99921560)

[4.5.3 Conclusion 10](#_Toc99921561)

[4.6. Multivariate Adaptive Regression Splines (MARS) 10](#_Toc99921562)

[4.6.1 Implementation 10](#_Toc99921563)

[4.6.2 Results 11](#_Toc99921564)

[4.6.3 Conclusion 11](#_Toc99921565)

[4.7 Fostering Positive Seller Impression Through Machine Learning 11](#_Toc99921566)

[5. Understanding Customer Value 11](#_Toc99921567)

[5.1 Introduction to the Analytics Solutions 11](#_Toc99921568)

[5.2.1 Time Series Forecasting 11](#_Toc99921569)

[5.2.2 Random Forest 12](#_Toc99921570)

[5.2.3 Logistic Regression 12](#_Toc99921571)

[5.2.4 Multivariate Adaptive Regression Splines 12](#_Toc99921572)

[5.2 Measures and Targets for CLV Forecasting and Clustering 12](#_Toc99921573)

[5.3 Time Series Forecasting 12](#_Toc99921574)

[5.3.1 Data Preparation 12](#_Toc99921575)

[5.3.2 Hyperparameter Tuning for ARIMA 12](#_Toc99921576)

[5.3.3 Results 13](#_Toc99921577)

[5.3.3 Accuracy Comparisons within Datasets 14](#_Toc99921578)

[5.3.4 Accuracy Comparisons across Datasets 14](#_Toc99921579)

[5.4 Data Preparation 14](#_Toc99921580)

[5.4.1 Calculating RFM per Customer 14](#_Toc99921581)

[5.4.2 Normalising RFM Scores and Obtaining CLV 15](#_Toc99921582)

[5.4.3 Segmenting Customers based on CLV Using K-Means Clustering 15](#_Toc99921583)

[5.5. Datasets Used 15](#_Toc99921584)

[5.5.1 Importance of Balancing to Solving our Business Problem 15](#_Toc99921585)

[5.6 Random Forest 16](#_Toc99921586)

[5.6.1 Variable Importance 16](#_Toc99921587)

[5.6.2 Accuracies 16](#_Toc99921588)

[5.6.3 Hyperparameter Tuning 16](#_Toc99921589)

[5.6.4 Conclusion 16](#_Toc99921590)

[5.7 Logistic Regression 17](#_Toc99921591)

[5.7.1 Variable Importance 17](#_Toc99921592)

[5.7.2 Accuracies 18](#_Toc99921593)

[5.7.2 Conclusion 18](#_Toc99921594)

[5.8 Multivariate Adaptive Regression Splines (MARS) 18](#_Toc99921595)

[5.8.1 Parameter Tuning 18](#_Toc99921596)

[5.8.2 Variable Importance 18](#_Toc99921597)

[5.8.3 Accuracies 19](#_Toc99921598)

[5.8.4 Conclusion 19](#_Toc99921599)

[5.9 Optimising Customer Lifetime Value Through Machine Learning 19](#_Toc99921600)

[6. Conclusion 20](#_Toc99921601)

[References 21](#_Toc99921602)

[Appendix A: Association Rules 25](#_Toc99921603)

[A.1 Top 10 Association Rules in Terms of Lift Value (UCI Dataset) 25](#_Toc99921604)

[A.2 Bottom 10 Association Rules in Terms of Lift Value (UCI Dataset) 25](#_Toc99921605)

[A.3 Association Rules between Products (Olist Dataset) 26](#_Toc99921606)

[A.4 Association Rules between Sellers (Olist Dataset) 27](#_Toc99921607)

[Appendix B: Linear Regression 28](#_Toc99921608)

[B.1 Derived Variables 28](#_Toc99921609)

[B.2 Diagnostic Plots 28](#_Toc99921610)

[Appendix C: Quantile Regression 29](#_Toc99921611)

[C.1 Quantile Regression Result on the Original Variables 29](#_Toc99921612)

[C.2 Quantile Regression Result with Additional Derivative Variables 30](#_Toc99921613)

[Appendix D: Time Series 31](#_Toc99921614)

[D.1 Exponential Smoothing Methods 31](#_Toc99921615)

[D.2 Autoregressive Integrated Moving Average (ARIMA) Method 32](#_Toc99921616)

[D.3 Plots of Actual and Forecasted Data 35](#_Toc99921617)

[Appendix E: Random Forest 42](#_Toc99921618)

[E.1 Bootstrap Aggregating (“bagging”) 42](#_Toc99921619)

[E.2 Random Subset Feature selection (“RSF”) 42](#_Toc99921620)

[E.3 Random Forest Variable Importance Plots 42](#_Toc99921621)

[E.4 Random Forest Error Rates against Number of Trees Plots 43](#_Toc99921622)

[E.5 Random Forest Hyperparameter Tuning 44](#_Toc99921623)

[Appendix F: K-Means 45](#_Toc99921624)

[F.1 K-Means Clustering Process 45](#_Toc99921625)

[Appendix G: Logistic Regression 47](#_Toc99921626)

[G.1 Multinomial Confidence Intervals for UCI Original Trainset 47](#_Toc99921627)

[G.2 Multinomial Confidence Intervals for Olist Original Trainset 48](#_Toc99921628)

[G.3 P-Value of Variables for UCI Original Trainset 49](#_Toc99921629)

[G.4 P-Value of Variables for Olist Original Trainset 49](#_Toc99921630)

[Appendix H: MARS 51](#_Toc99921631)

[H.1 Results of MARS Model with degree = 2 - Olist Original Trainset 51](#_Toc99921632)

[H.2 Results of MARS Model with degree = 1 - Olist Original Trainset 51](#_Toc99921633)

[H.3 Results of MARS Model with degree = 1 - Olist Balanced Trainset 51](#_Toc99921634)

[H.4 Results of MARS Model with degree = 2 - Olist Balanced Trainset 53](#_Toc99921635)

[H.5 Results of MARS Model with degree = 1 - UCI Original Trainset 54](#_Toc99921636)

[H.6 Results of MARS Model with degree = 2 - UCI Original Trainset 54](#_Toc99921637)

[H.7 Results of MARS Model with degree = 1 - UCI Balanced Trainset 55](#_Toc99921638)

[H.8 Results of MARS Model with degree = 2 - UCI Balanced Trainset 55](#_Toc99921639)

[H.9 Variable Importance generated by MARS Model - UCI Original Trainset 56](#_Toc99921640)

[H.10 Variable Importance generated by MARS Model - UCI Balanced Trainset 56](#_Toc99921641)

[H.11 Variable Importance generated by MARS Model - Olist Original Trainset 57](#_Toc99921642)

[H.12 Variable Importance generated by MARS Model - Olist Balanced Trainset 57](#_Toc99921643)

# EXECUTIVE SUMMARY

#### Problem Statement

There is currently a rise in the number of companies adopting the Direct-to-Consumer (DTC) model rather than selling their goods on an e-commerce marketplace. This causes e-commerce platforms to have greater difficulty in acquiring retailers.

#### Our Solution and Results

This report discusses the ways e-commerce platforms can offer analytics-based insights and benefits to attract sellers onto their platform rather than adopting a DTC model.

#### Findings around the Business Problem

The Pareto principle applies in customer segmentation by Customer Lifetime Value, where a large majority of customers provide low value in contrast to a few, high-value customers.

#### Findings around our Analytics Models

Association rule mining is more effective on the UCI dataset (generating more association rules) as compared to on the Olist dataset. This is due to the UCI dataset having more transactional data as compared to the Olist dataset.

From the association rules derived, it is observed that products that are more similar tend to have greater lift values, hence greater complementary effects.

Customer Review Score prediction with Quantile Regression yields the best accuracy at 50th percentile of review scores. Additionally, the value of Quantile Regression comes in the form of suggestions that can be given to our sellers on how they can improve their products at each percentile of review scores. For MARS, the model performed better on the derived dataset with degree 1 than degree 2, as well as better than both degree 1 and 2 on the dataset without derived variables. MARS had the overall best accuracy for our business opportunity.

In time-series forecasting for Customer Lifetime Value, Autoregressive Integrated Moving Average (ARIMA) performs better than any other forecasting method that makes predictions by modelling trends and/or seasonality. This includes Simple Exponential Smoothing, Holt’s Method, and Winter’s Method.

In time-series forecasting for Customer Lifetime Value, performing manual hyperparameter tuning for ARIMA results in forecasts that perform at least as good as their automated tuning counterparts.

In cluster prediction for customers via Customer Lifetime Value and the Recency, Frequency, and Monetary method of computation, it is important to balance the data to obtain higher prediction accuracies on high-value customers.

In Random Forest, the default hyperparameters provided by R have acceptable performance, eliminating the need for manual tuning as they provide minimal returns.

In Logistic Regression, the default parameters provided by R were sufficient to generate a high accuracy model while showing the variable importance of each level.

In MARS (Categorical), degree 2 was specified to have the highest GRSq and accuracy. Individual breakdowns of variable importance were sufficient for our business needs.

# 1. The Business Problem / Opportunity Statement

## 1.1. The Opportunity Statement

The recent rise in adoption of the Direct-to-Consumer (DTC) model rather than selling on e-commerce marketplaces causes e-commerce platforms to lose out on both potential and current retailers. This has exacerbated in recent years due to the Covid-19 pandemic, where DTC sales growth was 45.5% from 2019 to 2020, and US digital DTC sales are expected to reach $175 billion by 2023 (Geldman, 2021).

To remain competitive and tackle the growing difficulty in acquiring retailers, e-commerce marketplaces need to leverage on their greatest competitive advantage - their wealth of data. Amazon alone has two billion US visits per month, followed by eBay and Walmart.com with 688.9 million and 388.8 million US visits respectively (Geldman, 2021). By sharing data-driven insights with sellers, marketplace companies can offer sellers a unique selling point to attract them onboard.

This benefits the marketplace companies by increasing their market share and preventing competing models like DTC from eating away at their margins. More sellers means not only more revenue, but more data would also be generated. This generates a positive reinforcing loop that enhances the impact and accuracy of future predictions. These data-driven insights can also serve as a potential source of extra revenue if sellers are charged for them.

In this report, we aim to produce analytics-driven solutions for e-commerce marketplaces to provide to sellers, which focus on countering the benefits of DTCs as opposed to e-commerce marketplaces. This will ultimately allow us to assist e-commerce marketplaces in retaining and acquiring sellers onto their platforms.

## 1.2. Justification of Business Problem

Our business problem has two motivating factors - the rapid adoption of ecommerce in recent times, and the growing feasibility and appeal of going Direct-to-Consumer (DTC) as compared to going on marketplaces.

### 1.2.1 Rise of E-Commerce

In 2020 the global ecommerce industry saw an astounding 24% spike in sales compared to the previous year (Straight, 2021), producing sales revenue of over $4.2 trillion. (Verdon, 2021) This bullish trend is not expected to stop anytime soon, as it is forecasted that retail ecommerce sales will continue to grow by double digits through 2023, reflecting a near 10% increase in retail market share by volume (Lebow, 2021)

.

The drivers of this explosive growth are self-evident. Catapulting five-years of growth into a single quarter to accelerate the shift from offline to online for businesses and consumers, COVID-19 is perhaps the most monumental black swan event of the 2020s, and its impacts on the way we live, work, and play cannot be understated. A paradigm shift emerged in consumer behaviour as people around the world are locked down in their homes and forced to do their shopping through online channels, and business models are pressured to quickly follow in keeping up with the consumers (McKinsey & Co., 2020). Of course, other factors such as technological advancements and perceived price advantages have also contributed to his growth.

Nonetheless, with the increasing allure of ecommerce also comes increased competition. Here, we take on the perspective of marketplace retailers as we discuss the emergence of a direct competitor - the Direct-to-Consumer (DTC) model. Understanding our competition is important as it can serve as a guide on what specific benefits marketplace retailers should start to offer in response to this DTC threat. This link will become evident in the Approach section in Section 1.3.

### 1.2.2 Emergence of DTC Model

Direct-to-Consumer (DTC) refers to a distribution model where companies sell directly to their customers, i.e. bypassing any third-party retailers, wholesalers, or any other middlemen. Its emergence has been empowered by two factors, its inherent appeal, as well as its increasing feasibility over the years with the advancement of technology. We will discuss these two factors in order.

#### 1.2.2.1 The Appeal of DTC - Expanding and Enhancing Seller’s Presence

DTC gives brands the flexibility to expand and enhance their presence by controlling all aspects of consumer interactions with the brand without an intermediary, forming a crucial part of omnichannel marketing. (Goulart, 2020)

#### 1.2.2.2 The Appeal of DTC - Greater Opportunities to Build Positive Customer Relationships

Since DTC allows each brand direct access to consumers through their own channels, it can build better customer relationships by connecting directly with them and offering content that resonates with them best. (eDesk, 2021)

#### 1.2.2.3 The Appeal of DTC - Increased Margins

DTC allows for greater margins by cutting out middlemen. Studies have shown that successful DTC companies have a gross margin of 50-85%, as removing the need for distribution partners ends up saving costs for the company tremendously. (Shruthi, 2021)

#### 1.2.2.4 The increasing feasibility of managing a DTC model

Regardless, having innate appeal is useless if brands do not have a practical way of implementing a DTC model. This brings us to the second part of the equation - technological advancements that make DTC much more feasible .

Technological advancements have lowered the barriers of entry in adopting a DTC model. With the help of Artificial Intelligence such as chatbots and machine learning, brands are also able to easily manage and improve their DTC model. This has led to a democratisation of ecommerce, with Shopify (Sularia, 2021) and SEKO (DC Velocity, 2022) leading the charge in shifting power away from ecommerce marketplace retailers.

#### 1.2.2.5 Case studies of brands that successfully adopted the DTC model, and DNVBs

Nike’s success story with adopting the DTC model is sure to drive up the appeal of DTC. Through the sportswear giant’s continuous focus on DTC growth, coupled with the recent Consumer Direct Acceleration strategy to speed up prioritisation of DTC, Nike’s DTC sales grew from $2.9 billion in 2011 to $12.4 billion by the end of Nike’s fiscal 2020. (Salpini, 2021)

This growth can also be observed in DTC brands such as Allbirds and Glossier, who are currently valued at $1.7 billion and $1.8 billion respectively after starting out in the mid 2010s (Koss, 2022).

Today, the DTC model has gone on to inspire the Digitally Native Vertical Brand, or DNVB - a term to describe a brand that was born online and sells their own products on their own website, claiming full control over their customer experience, from factory to consumer (Big Commerce, 2022).

These trends present a tremendous opportunity for analytics to be applied. By leveraging the tons of data being generated through the daily transactions in the ecommerce space, data-driven insights could prove valuable in increasing revenue, user satisfaction, and market share. These, in turn, can help us analyse how marketplace retailers can provide value to sellers, and help us understand how marketplace retailers may remain resilient in the face of new competition (in the form of the DTC model).

## 1.3 Desired Business Outcomes

To concretise our focus areas and structure our approach, we will refer to the Justification of Business Problem - specifically, on the [appeal of DTC](#_heading=h.r1h7e9ewjicz) and how marketplace retailers can provide a substitute - or even better alternative - to those benefits.

### 1.3.1 Cost-Effective Means to Expand and Enhance Seller Presence

Tackling [1.2.2.1](#_heading=h.r1h7e9ewjicz), we can offer a similar benefit by recommending their products to new customers based on their purchasing patterns. We do this via association rules in [Section 3](#_heading=h.ab6dpgfo0as5), where we make use of association rule mining to identify underlying relations between different products and different sellers, thus understanding customers’ purchasing patterns. Using the results, we can recommend customers products based on the products they have in their carts. This helps brands increase their product visibility and expand their customer base as compared to selling on their own page.

Furthermore, online marketplaces can offer brands to personalise their stores within the marketplace to control their branding and experience, a strategy long adopted by Alibaba (Walk, 2021).

### 1.3.2 Fostering Positive Seller Impressions and Customer Relations

Moving on to [1.2.2.2](#_heading=h.elschgng5qg0), we offer an alternative method of building customer rapport by offering brands a way to predict how customers would rate their products, and make recommendations to them accordingly. We do this using machine learning techniques in [Section 4](#_heading=h.9yxs1m1s7z6o) to predict customer ratings, so as to allow companies to have an estimate of their ratings and suggest factors they can improve on by observing the key predictors.

### 1.3.3 Business Strategy Excellence through Understanding of Customer Lifetime Value (CLV) Trends and Drivers

Lastly, for [1.2.2.3](#_heading=h.yzbg4wsp9wny), we will offer a way for brands to understand their customer lifetime value (CLV) in terms of its temporal (trend) and spatial (important variables) characteristics. CLV is important as it costs less to keep existing customers than it does to acquire new ones, so increasing the value of existing customers is a great way to drive growth (Qualtrics, 2022). We do this through time-series forecasting on CLV as well as machine learning techniques on customer segment prediction in [Section 5](#_heading=h.d7r7w8ar6rdy). This will provide DTC companies with insights on how their CLV is expected to perform in the future and what they can do to increase it, informing their business strategy in managing CLV.

# 2. Data Selection and Cleaning

Data cleaning is an integral part of the analytics process (Thomas, 2019). To prepare for our modelling, we found and cleaned two separate datasets.

## 2.1 Data Cleaning: UCI eCommerce

This first dataset was obtained from the UC Irvine Machine Learning Repository and can be downloaded here: <https://archive.ics.uci.edu/ml/datasets/online+retail>. The steps described in this section can be found in ‘*Data Cleaning for UCI.ipynb*’.

### 2.1.1 Handling Missing Values

24% of transactions had missing CustomerIDs. Since CustomerID is an essential column to us (see Section 5), we remove these rows. As for UnitPrice, as there was no indication that products could be sold for free, we imputed rows where UnitPrice = 0 with the average price for that product.

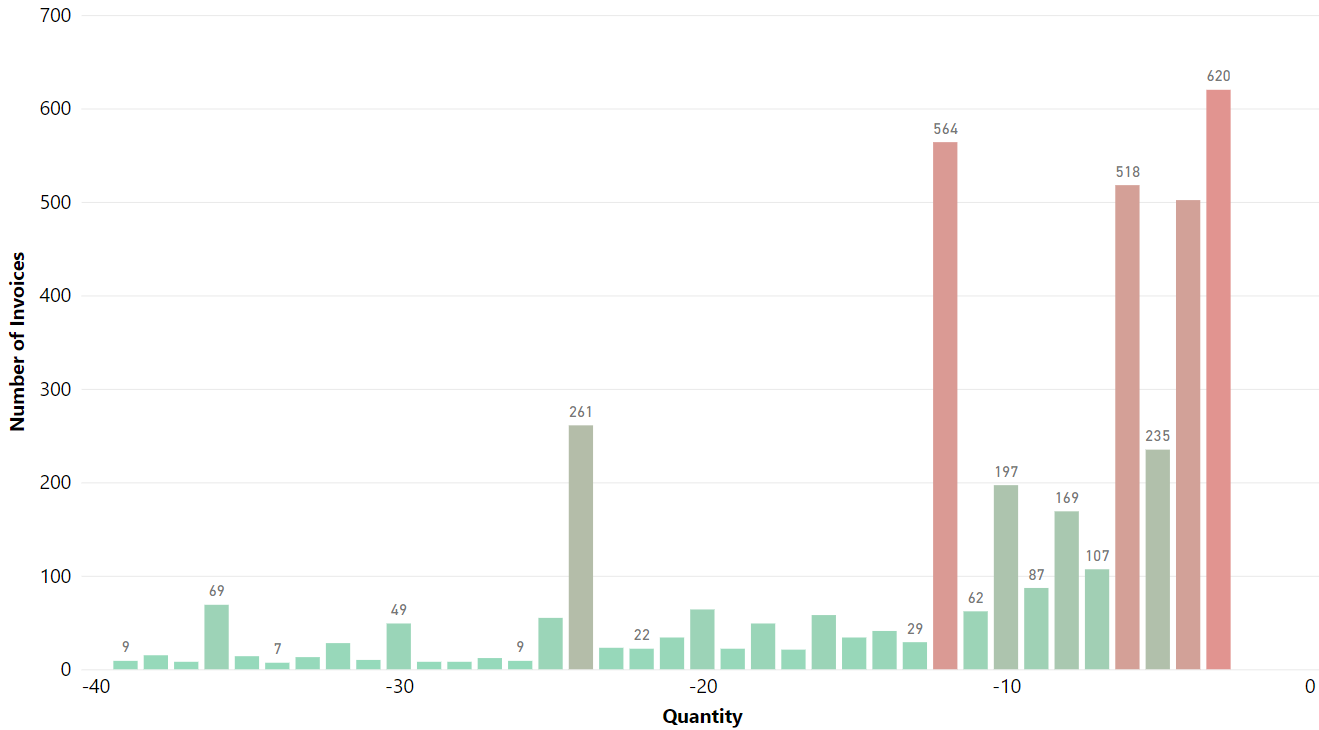
### 2.1.2 Filtering Non-Purchase Transactions via InvoiceNo and Quantity

According to the data dictionary, InvoiceNo should be a ‘6-digit integral number’, and those ‘start(ing) with the letter 'C' were cancelled’. We found 3839 rows (1%) not in integer format, and 3 rows that started with ‘A’:



We were unable to trace these bad debts to their source, and thus removed them.

Additionally, 3% of transactions had negative quantities (depicted below, left), most of which were due to cancellations as discussed in InvoiceNo. 13% of these transactions, however, were related to inventory upkeep as seen from their Description (depicted below, right). Since we are only concerned with actual sales data, we will remove these transactions related to upkeep.

### 2.1.3 Handling Cancelled Transactions

For cancelled transactions, it is important to match them to their corresponding order transaction where possible, so as to derive the net quantity transacted. A product/customer with many initial orders but ultimately have had most of them cancelled/refunded does not actually bring much value to a store.

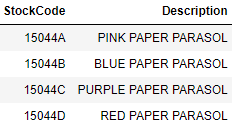
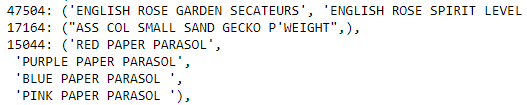




We can make logical inferences in matching cancelled transactions with their original order through their StockCodes and CustomerID. By doing so, and by validating our match using InvoiceDate (cancellation date must be after order date) and Quantity (quantity cancelled must be less than quantity ordered), we managed to trace 89% of cancellations to their original order. We speculate that the invalid matches are due to cancellations of products ordered outside of the dataset’s timeframe of ‘01/12/2010 and 09/12/2011’.

### 2.1.4 Deriving a new ‘Product Variations’ column from StockCode

According to the data dictionary, StockCode should be a ‘5-digit integral number’. Some products had 6 digits instead where an alphabet was appended to the back (depicted below, left), and stripping these StockCodes to their base 5-digit form reveals that this alphabet is to indicate a product variation (depicted below, right).

This allowed us to derive a new column, ProductVariations, for each product.

|  |  |
| --- | --- |
| 2.2 Data Cleaning: Olist eCommerce The Olist dataset was downloaded from Kaggle, containing Brazilian e-Commerce Marketplace orders data by Olist which can be downloaded from <https://www.kaggle.com/olistbr/brazilian-ecommerce>.  Olist connects small businesses all over Brazil to their customers, and the dataset contain information of about 100,000 orders from 2016-2018 made at the marketplaces in Brazil. |  |

The data is divided into multiple datasets which organizes the data into multiple dimensions: order status, price, payment and freight performance, location, etc. The schema above portrays the relation between datasets.

### 2.2.1 Selecting “delivered” Status

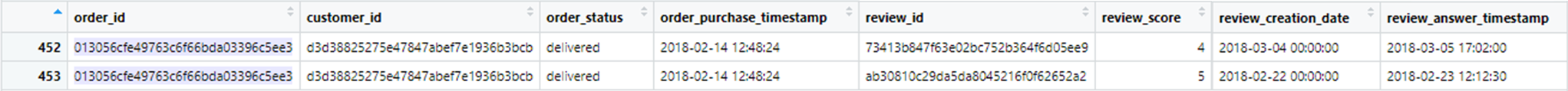
The olist\_orders\_dataset.csv contains orders of multiple different status including “delivered”, ”canceled”, ”processing”, “unavailable”, etc. Our team has decided to only analyze completed and “delivered” orders. Therefore, the data was subset to only include orders with “delivered” status.

### 2.2.2 Merging Data

The raw data consist of 9 datasets (the 8 datasets portrayed in the schema above and 1 product category translation dataset). As the data are stored in multiple files, data merging is necessary to enable us to analyze them collectively.

### 2.2.3 Removing Duplicates

Some of the data merging processes cause duplicates to happen. For example, a person buying a particular product might give a review and update the review sometime later after trying out the product. As a result of the multiple reviews on a particular order\_id, a particular row in the orders dataset can be duplicated twice/more when orders dataset and reviews dataset are merged (depicted below). These duplicates need to be removed, selecting only the most recent review in the final merged data. Similar occurrences happen in the process of merging other datasets.



### 2.2.4 Converting Time Data to the Right Format

When the data is imported into R, the date-time data is stored as a character. Converting the date-time data to POSIXct class allows us to do time processing, such as finding the time difference between two data, making conditions based on time, etc.

# 3. Expanding and Enhancing Seller Presence

## 3.1. Introduction to the Analytics Solution: Association Rules

Association Rule mining is a data mining technique used to find interesting associations and relationships among items in a large dataset (D, 2018). With e-commerce platforms’ large transactional databases, we aim to make use of association rule mining to help these platforms to expand and enhance seller presence..

An association rule is made up of two parts: an antecedent (if), item(s) found in the dataset, and a consequent (then), item(s) found in combination with the antecedent in the dataset.

## 3.2 Measures for Association Rules

There are three measures of effectiveness of association rules: support, confidence and lift.

1. Support refers to the frequency of occurrence of the antecedent of an association rule and is a measure of the “applicability” of the rule, that is how often the rule can be applied.
2. Confidence refers to the number of times the association rule turns out to be true, which indicates the strength of the association rule.
3. Lift is a measure of the importance of an association rule. A lift value greater than 1 indicates a positive dependence between the antecedent and consequent (complementary effect) whereas a lift value smaller than 1 indicates a negative dependence (substitution effect) (Baesens, 2017).

These three measures are used to mine and filter meaningful association rules.

## 3.3. Association Rules on UCI Dataset

We applied association rules to find meaningful associations between different products sold based on unique transactions.

### 3.3.1 Data Preparation

We created a new column, ‘product’, which is a string consisting of the description as well as the unique stock code of the product.

We then converted the UCI transaction dataset (originally long format) into a wide format dataframe, with rows representing each transaction based on ‘InvoiceNo’ and columns representing each product based on ‘product’. For the columns, we assigned the value ‘1’ if the transaction contains the product. Otherwise, we assign the value ‘0’.

Afterwards, we converted the wide format dataframe into a transaction object for association rule mining.

### 3.3.2 Results

A total of 748 sets of association rules which satisfied the minimum support (0.01) and minimum confidence (0.10) conditions and lift greater than 1 were identified.

### 3.3.3 Insights

The association rules with higher lift values (Refer to Appendix A.1) tend to be made up of a variation of the same product (e.g. different colours).

On the other hand, association rules with lower lift values (Refer to Appendix A.2) tend to be made up of different types of product.

## 3.4. Association Rules on Olist Dataset

We applied association rules to find meaningful associations between different products sold and between sellers based on purchases made by customers within the same quarter.

### 3.4.1 Data Preparation

We created two new columns, ‘customerUID\_year/quarter’ and ‘product’. ‘customerUID\_year/quarter’ is a string consisting of the unique customer id as well as the quarter/year the purchase was made and ‘product’ is a string consisting of the product category, seller id and product id of the product.

To apply association rule mining to find associations between products, we selected only 2 variables, ‘customerUID\_year/quarter’ and ‘product’, and converted the dataframe into a transaction object.

For finding associations between sellers, we did the same but selected ‘seller\_id’ instead of ‘product\_id’.

### 3.4.2 Results

Association rules between products

A total of 29 sets of association rules which satisfied the minimum support (0.00005) and minimum confidence (0.10) conditions were identified. All 29 of them have lift greater than 1 as well.

Association rules between sellers

A total of 15 sets of association rules which satisfied the minimum support (0.00005) and minimum confidence (0.10) conditions were identified. All 15 of them have lift greater than 1 as well.

Rationale for low minimum support

Despite the large amount of transactional data (107823), most of these transactions only consist of one item and only 15% of the transactions are of repeat customers. Hence, the support values of association rules existing within the dataset are generally very low, so we set the minimum support to be 0.00005 to capture the relatively more frequent association rules.

### 3.4.3 Insights

From the association rules mined (Refer to Appendix A.3), we can tell that consumers are likely to purchase certain products of the same category together (and likely of the same seller). There are also association rules between products sold by different sellers.

However, the association rules between sellers (Refer to Appendix A.4) have lower lift values as compared to that of those between products (Refer to Appendix A.3), which indicates a weaker complementary effect.

## 3.5. Expanding and Enhancing Seller Presence Through Association Rules

Enhancing seller presence

The use of association rules mining allows e-commerce marketplaces to find associations between products sold by the same seller with strong complementary effects (very high lift values). These association rules can be applied in designing more effective recommendation systems to enhance a seller’s presence by recommending buyers products from the same seller, which they are highly likely to purchase, when products are added into their carts.

Expanding seller presence

The same can be done to expand a seller’s presence when we apply the use of association rules mining to find strong dependencies between products sold by different sellers. Even though these association rules tend to have lower lift values, the lift values are still greater than 1. This shows that the association rules can still be applied to boost the sales of different sellers when recommendations are made based on them. Hence, e-commerce marketplaces can apply association rules mining to recommend products from other sellers, based on the seller of the products added to the customers’ carts.

UCI vs Olist

The association rule mining technique is more effective on the UCI dataset (394271 rows) due to a larger number of records in the dataset as compared to the Olist dataset (107823 rows). This resulted in a greater number of association rules mined from the UCI dataset as compared to Olist, even when the minimum support is set higher, generating more information on customer behaviour. Hence, in reality where marketplaces are likely to have larger transactional databases than DTCs, association rule mining will be more effective for marketplaces to understand customer purchasing patterns.

# 4. Building Positive Customer Relationships

## 4.1. Introduction to the Analytics Solutions

### 4.1.1 Linear Regression

Linear Regression is the supervised machine learning model in which the model finds the best fit linear line between the independent and dependent variable. In our case, we used the variables in the dataset (dependent variables) to predict customer review score (independent variable).

### 4.1.2 Quantile Regression

Quantile regression is an extension of linear regression used when the assumptions of linear regression are not fulfilled (i.e. linearity, homoscedasticity, independence, or normality). Unlike regular linear regression which uses the method of least squares to calculate the conditional mean of the target across different values of the features, quantile regression estimates the conditional median of the target (Dye, 2020).

### 4.1.3 Multivariate Adaptive Regression Splines (MARS)

When interaction between metrics is non-linear, simple linear regression may not be able to give us a good approximation of outputs given the inputs. This is true for our dataset as mentioned 4.4.3. Instead, we can use MARS which is an ensemble of linear functions joined together by one or more hinge functions (Dobilas, 2020).

## 4.2. Measures and Targets for Review Score Predictions

We aim to build models with a high accuracy, especially in predicting the low review scores. This will help us recommend how our sellers can improve their performance. Before we dive into the different machine learning techniques used to predict review\_score, we first created a function to calculate the accuracy of our predictive models. We will use this in Sections 4.4, 4.5 and 4.6 where each model is run on the testset to generate predictions, and the function calculates the accuracy by comparing the predictions with actual review scores in the testset.

|  |  |
| --- | --- |
| The review score is a discrete numeric variable taking integer values of 1, 2, 3, 4, or 5. However, the review score predictions generated by the models are continuous data. Therefore, the prediction needs to be classified to enable 1-to-1 comparison with the actual review score. The classification is done as shown on the right. | Prediction score <= 1.4 is given a score of 1  1.4 < Prediction score <= 2.4 is given a score of 2  2.4 < Prediction score <= 3.4 is given a score of 3  3.4 < Prediction score <= 4.4 is given a score of 4  Prediction score > 4.4 is given a score of 5 |

## 4.3. Data Preparation

For the purpose of modelling, a number of variables in the dataset were dropped as they were perceived to have no relevance in predicting the review score. These are variables like order\_id, customer\_id, order\_status, review\_creation\_date, etc. The final dataset contains these 9 variables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| review\_score | payment\_type | payment\_installments | payment\_value | price |
| freight\_value | product\_name\_lenght | product\_description\_lenght | product\_photos\_qty |  |

The dataset was then split into train-test datasets with a 70/30 ratio.

## 4.4. Linear Regression

### 4.4.1 Implementation

We trained our linear regression model on the trainset (‘orders\_all\_1’) using all the variables in the dataset (other than review score). Next, we tested our model on the testset, calculated the accuracy and retrieved various results.

Additionally, we added several variables, including payment\_type and derived variables to our dataset, then retrained the linear regression model. The new dataset (‘orders\_all\_2’) includes the variables seen in Appendix B.1.

We then split the data into trainset and testset, trained the new model on the new trainset and got the results.

### 4.4.2 Results

|  |  |
| --- | --- |
| We used the function in [Section 4.2.](#_heading=h.s1rkubs7w5it) to calculate the accuracy of both models as shown in the table below  Evidently, the linear regression model trained on the second dataset with additional variables has a higher testset accuracy of 0.266 compared to 0.191 for the model trained on the initial dataset. |  |

The significant variables in the second model with p-value < 0.001 include payment\_typevoucher, payment\_installments, payment\_value, freight\_value, product\_name\_lenght, product\_description\_lenght, del\_time, Late1, total\_price and freight\_ratio. Some of these variables such as product name length and product description length can be used to make recommendations to individual sellers on how to make their products more favourable.

### 4.4.3 Conclusion

Linear regression can be effective as it is easier to implement, interpret and efficient to train than other models.

However, linear regression has four assumptions - linearity, homoscedasticity, independence, and normality. After plotting the four diagnostic plots as shown in Appendix B.2, we realise that linear regression is not apt for our dataset.

In the Residuals vs Fitted plot, the line is far from a horizontal line and with distinct patterns, indicating a non-linear relationship. For the Normal Q-Q plot, the residuals fluctuate about the straight dashed line, which shows that the residuals are not normally distributed. In the Scale-Location plot, the line drawn is sloping down (not horizontal), and definitely not with evenly spread points. This indicates that our dataset has a heteroscedasticity problem. For the final Residuals vs Leverage plot which is used to identify influential cases, observations with standardised residuals greater than 3 are possible outliers (Hastie et al., 2013). In our plot, there are a significant number of points that exceed 3 standard deviations. Additionally, there are high leverage points in the data, with a leverage statistic below 2(p + 1)/n = 4/200 = 0.02. These indicate the presence of many outliers in our dataset (Kassambara, 2018).

Since all four of the linear regression assumptions are not met, linear regression is not effective for our dataset and we will move on to other machine learning techniques - quantile regression and MARS.

## 4.5. Quantile Regression

### 4.5.1 Implementation

Quantile regression was done on the trainset (‘orders\_all\_1’) across multiple tau values, namely 0.1, 0.25, 0.5, 0.75, and 0.9. The models were run on the testset to generate prediction, calculate the accuracy, and retrieve the coefficients and p-values. This was done to compare how the accuracy of each model and the significance of each variable varies across quantiles.

Additionally, quantile regression was done on another trainset (‘orders\_all\_2.2’) which were modifications of ‘orders\_all\_2’ trainset. The modifications consist of removal of several variables that were linearly independent (‘est\_del\_time’ and ‘del\_time’), causing singular matrix error when quantile regression was run.

### 4.5.2 Results

The results of quantile regression on ‘orders\_all\_1’ trainset across the different tau values were combined into a dataframe as shown in Appendix C.1.

The accuracy calculated from the function in [Section 4.2.](#_heading=h.s1rkubs7w5it) is shown on the first row. Subsequent rows contain the coefficients of each variable across multiple quantiles. The row below each variable coefficient with empty variable name contain the p-values of each variable across the different quantiles.

There program showed a singular matrix error when quantile regression was done with 0.75 and 0.9 tau values, indicating that the data structure is not suitable for these quantiles. As a result, no values are shown in the last 2 columns.

As seen in the table, the accuracy tends to increase with higher quantiles, with the max accuracy of 57.6% achieved at quantile regression for the 50th percentile. Some variables were significant at a certain quantile but not for the others as indicated by their p-value below 0.05. For example, the payment\_typedebit\_card is significant at the 10th percentile but not at the 25th and 50th percentile, the price variable is important only at the 25th percentile, and so on.

The result of quantile regression on ‘orders\_all\_2.2’ trainset is shown in Appendix C.2.

Again, the program showed a singular matrix error at 0.75 and 0.9 quantiles, indicating the data structure is not suitable for these quantiles.

With additional variables, the highest accuracy decreased to 56.27% at 50th percentile. The accuracy at 10th percentile increased to 10.5% while the accuracy at 20th percentile increased to 21.87%. A number of variables were shown to be significant only in the 10th percentile of review scores. These variables include payment\_typevoucher, product\_name\_lenght, product\_description\_lenght, product\_photos\_qty, total\_price, freight\_ratio, and purchase\_day\_of\_week. These variables can be used to make recommendations to individual sellers on how to make their low-rating products more favourable to the consumers.

The payment\_value variable is significant across all quantiles, however, the significance seems to decrease from the lower to higher quantiles. The Late variable has a p-value of 0 across all quantiles, indicating that it is significant in predicting review scores in all quantiles.

### 4.5.3 Conclusion

The result of quantile regression has a higher predictive accuracy (0.576) compared to linear regression (0.266). This resulted from the robustness of quantile regression on the type of distribution. Minimizing the absolute residuals will work well no matter how skewed the data is.

The higher accuracy of quantile regression at the 50th percentile likely arose from the skewed distribution of review scores where there is a larger percentage of review score 5 compared to review scores 1 to 4. Additionally, quantile regression describes the relationship between variables at different points in the conditional distribution of the outcome variable. As a result, we are able to find out which variables are important for predicting review scores at each quantile of the review score distribution. When the additional derivative variables were added, most of them were significant predictors of review scores at the 10th percentile, which could possibly explain the improved accuracy at this percentile.

## 4.6. Multivariate Adaptive Regression Splines (MARS)

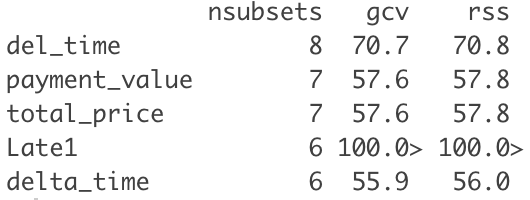
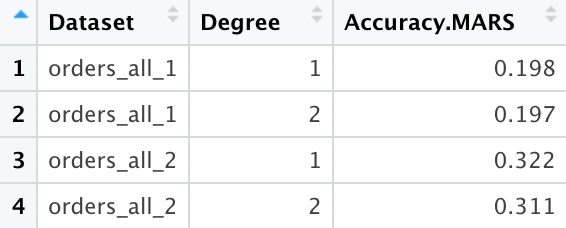
### 4.6.1 Implementation

We trained our MARS models on the original trainset (‘orders\_all\_1’), followed by the trainset (‘orders\_all\_2’) with derived variables obtained in [Section 4.4.1](#_heading=h.pa8kji4ukoax). To find the model with the best fit, we trained 4 different models - MARS Degree-1 and MARS Degree-2 for both trainsets.

For all 4 models, we included all the variables in the trainset (other than review score) when training. We then tested the models on the respective testsets and calculated the accuracy of each model using the function from [Section 4.3](#_heading=h.s1rkubs7w5it).

### 4.6.2 Results

Below are the testset errors for the 4 MARS models, where the first two entries are for the models trained on the original trainset with degrees 1 and 2 respectively, while the last two entries are for the models trained on the trainset with derived variables with degrees 1 and 2 respectively.



The model with the best performance is the model trained on the second dataset with degree 1, with an accuracy rate of 0.322.

### 4.6.3 Conclusion

Our chosen MARS model has a higher predictive accuracy (0.322) than our linear regression model (0.266). This is expected as our dataset does not meet the four assumptions of linear regression, and MARS is able to account for the different linear functions present in the dataset.

## 4.7 Fostering Positive Seller Impression Through Machine Learning

We identified 2 main objectives in our opportunity: generating accurate prediction and providing insights for product improvement.

[Linear Regression](#_heading=h.yfdrzdfz3d3u) is simple and easy to implement. However, the 4 assumptions of linear regression were not satisfied by our dataset, rendering it inaccurate and unsuitable for our purpose. [Quantile Regression](#_heading=h.e30lzk18qx8) has the best predictive accuracy among all models at 0.5 tau, but has lower accuracy for other tau values. Despite having the greatest accuracy, we do not know and cannot assume the tau value at the time of prediction, making it unsuitable for predicting review scores. However, the possibility to analyse which variables are significant at each quantile enables us to provide suggestions on how individual sellers can improve their customer review scores, leading to better relationships with customers. Hence, quantile regression is suitable for our second objective.

Our [MARS](#_heading=h.7ms9bbhv86um) model from Section 4.6 has a predictive accuracy of 0.322. Although this value is not very high, review score is a complex metric influenced by many factors, including those not in the dataset (e.g. damaged item, received product differs from pictures, similar to pictures but buyer decided he/she doesn't like it etc.). Some of these factors cannot be measured through the e-commerce platform as the product is usually shipped from the seller directly to the buyer and is not inspected by the e-commerce company. Hence, we continue to believe that our MARS model is the most apt for generating accurate review score predictions (out of the 3 models) given what can be observed and measured on e-commerce platforms.

# 5. Understanding Customer Value

## 5.1 Introduction to the Analytics Solutions

### 5.2.1 Time Series Forecasting

A time series is a sequence of data points occurring in successive order across time (Scott, n.d.). A time series can have a trend, where the average value is not stationary over time. It can also have seasonality, where a consistent pattern repeats in the data yearly. We explore three different time series methods in this vein of trend + seasonality models, called Simple Exponential Smoothing (SES), Holt’s, and Winter’s, elaborated in technical detail in Appendix D.1. We also explore another time series model, called ARIMA, based on a statistical concept called autocorrelation, elaborated in technical detail in Appendix D.2.

### 5.2.2 Random Forest

Random forest is an ensemble machine learning technique that creates many decision trees and uses the majority prediction as its answer (Tin, n.d.). It is based on the ideas of bootstrap aggregating (“bagging/B”) and Random Subset Feature selection (“RSF”), and these are elaborated in Appendix E.1 and E.2 respectively.

### 5.2.3 Logistic Regression

Logistic Regression is a supervised learning method used for classification problems. Logistic Regression is used for binomial and multinomial categorical outcomes where multinomial is a simple extension of binary logistic regression. This method uses maximum likelihood estimation to evaluate the probability of the outcome falling into a category.

### 5.2.4 Multivariate Adaptive Regression Splines

We specified in [Section 4.6](#_heading=h.7ms9bbhv86um) that MARS is an ensemble of linear functions. However, MARS can also be used to handle categorical dependent variables as it contains an in-built generalised linear model (GLM) function. (Hastie & Tibshirani, n.d.) For multinomial dependent variables, MARS will build a standard response model (3 in the case of our analysis). MARS will then predict the response in the form of a probability that is estimated only with the single GLM model for the indicator column for that factor level (Milborrow, 2020) as MARS does not assume that there is a relationship between the levels. (TIBCO, n.d.) Therefore, the predicted row sums will not equate to one. For the purpose of our analysis, we will assign the cluster based on the highest response probability in the prediction.

## 5.2 Measures and Targets for CLV Forecasting and Clustering

For time-series forecasting, we will calculate how Customer Lifetime Value (CLV) changes over time on a monthly basis and attempt to predict the trend in the next quarter (three periods ahead). Forecast accuracies have three types of measures: scale-dependent errors (such as Mean Absolute Error MAE and Root Mean Squared Error RMSE), percentage errors (such as Mean Absolute Percentage Error MAPE), and scaled errors (such as Mean Absolute Scaled Error MASE ) (Tin, n.d.). We will use all three types to evaluate our forecasting models.

For the machine-learning based models, we will calculate CLV per customer, cluster them and determine the variables that are important in clustering customers. As this becomes a classification problem, we will use a confusion matrix to evaluate our model performance. A confusion matrix is an n x n table comparing the actual values with their values as predicted by the model. Values on the diagonal indicate correct predictions. We strive to maximise the number of correct predictions.

## 5.3 Time Series Forecasting

### 5.3.1 Data Preparation

To calculate average CLV in a given timeframe, we use the following simplified formula (Nithyakumar, 2022):

where Avg Value = Total Revenue / Total Purchases and Avg Frequency = No of Purchases / No of Unique Customers

We chose months as our timeframe and calculated CLV per month. We found that Olist had missing data for the month of November 2016. We hence removed the 3 2016 datapoints since time series forecasting requires continuous observations in a single time frame.

### 5.3.2 Hyperparameter Tuning for ARIMA

ARIMA has three parameters to supply: p, d, and q. These parameters can be tuned manually by visually inspecting various plots, or can be automatically chosen by R. We perform both, and provide an explanation for our parameter derivation in Appendix D.2.

### 5.3.3 Results

For each forecasting model, we provide their train and test set respectively, and highlight the rows with the lowest error for trainset (yellow) and testset (orange) for a given dataset. Refer to Section 5.2 for information about the error metrics.

For the UCI dataset, we used a different scaling factor for MASE computation since it did not have enough data to compute seasonal differences. (Hyndman, 2016)

Additionally, we brought in a third Kaggle dataset <https://www.kaggle.com/datasets/rohitsahoo/sales-forecasting> to perform Winter’s forecasting as it required at least 2 periods of observations, which neither Olist or UCI had.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RMSE | MAE | MAPE | MASE |
| SES (UCI) | 50.0 | 40.1 | 6.32 | 0.968 |
| 93.9 | 72.3 | 12.1 | 1.74 |
| Holt’s (UCI) | 52.5 | 45.7 | 7.41 | 1.10 |
| 134 | 113 | 18.5 | 2.71 |
| ARIMA (UCI)  (0,0,1) | 46.8 | 36.8 | 5.84 | 0.887 |
| 52.2 | 51.4 | 8.08 | 1.24 |
| Auto ARIMA (UCI)  (0,0,0) | 49.4 | 37.1 | 5.86 | 0.894 |
| 63.5 | 62.3 | 9.60 | 1.50 |
| SES (Olist) | 7.80 | 6.56 | 4.77 | 0.687 |
| 6.89 | 5.32 | 3.97 | 0.558 |
| Holt’s (Olist) | 9.21 | 7.67 | 5.51 | 0.803 |
| 9.57 | 7.91 | 5.87 | 0.829 |
| Auto/ARIMA (Olist)  (0,0,1) for both | 6.23 | 5.15 | 3.77 | 0.539 |
| 3.95 | 3.35 | 2.48 | 0.351 |
| Holt’s (Kaggle) | 194 | 128 | 25.10 | 0.958 |
| 97.43 | 79.09 | 16.87 | 0.590 |
| Winter’s (Kaggle) | 130 | 97.6 | 19.6 | 0.728 |
| 83.2 | 71.0 | 14.5 | 0.530 |
| ARIMA (Kaggle)  (1,0,1) | 103 | 77.9 | 17.2 | 0.582 |
| 59.6 | 52.4 | 10.8 | 0.391 |
| Auto ARIMA (Kaggle)  (1,0,0, 1,0,0) | 108 | 88.3 | 20.0 | 0.659 |
| 84.5 | 72.9 | 13.3 | 0.544 |

### 5.3.3 Accuracy Comparisons within Datasets

ARIMA consistently outperformed the trend and seasonality-based models (SES, Holt’s, Winter’s), having the lowest score in all error metrics for all datasets. Additionally, ARIMA models that have their p, d, and q values manually tuned (instead of relying on R’s auto.arima function) perform at least as good, if not better, than their Auto ARIMA counterparts. Comparing within the UCI and Kaggle dataset, ARIMA outperforms Auto ARIMA by roughly 5 points in RMSE. It was only on-par with Auto ARIMA in the Olist dataset, where both provided the same hyperparameters.

### 5.3.4 Accuracy Comparisons across Datasets

RMSE and MAE are scale-dependent, as such we can only look at MAPE, a percentage-dependent metric, and MASE, a scale-dependent metric. MASE is more suited to intermittent-demand series (Hyndman, n.d.), which suits our situation since certain products can have zero demand from time to time.

Between UCI and Olist, ARIMA on Olist performs better in both metrics. ARIMA on Olist performs similar to ARIMA on Kaggle in terms of MASE, even though ARIMA on Kaggle has a significant MAPE. Since we will use MASE as our error metric of choice, we can generally say that Olist and Kaggle performed better than UCI.

|  |  |
| --- | --- |
| **5.3.5 Conclusion**  Olist and Kaggle have data for longer periods than UCI. This, combined with their lower MASE, allows us to infer that having longer time-series data can generally result in better forecasting.  Plots of the actuals and forecasted data are shown for the best models for UCI and Olist. All plots can be found in Appendix D.3. |  |

|  |  |
| --- | --- |
| Forecasts are generally accurate across the board, with the actual value falling within the 80% confidence interval of the forecasted value.  The main limiting factor behind time-series forecasting is that it tells us *how* CLV changes over time, but not *what* causes CLV to change. While it can measure a seller’s success in managing CLV, it alone cannot inform business decisions. This motivates us to apply machine learning models to predict CLV and its key drivers. |  |

## 5.4 Data Preparation

The equation discussed in Section 5.3.1 is to calculate average CLV for an entire customer base in a given time period. Assigning CLVs to specific customers requires a different set of steps, which we describe in this section.

### 5.4.1 Calculating RFM per Customer

Calculating CLV requires us to first segment our customers in RFM - recency, frequency, and monetary. (Segal & Kvilhaug, n.d.)

A customer with a recent purchase is more likely to repeat a purchase than one who hasn’t purchased in a while.

A customer who frequents the store is also more likely to be a regular customer who brings higher value than one who rarely makes purchases, regardless of purchase value.

Lastly, the amount of money spent by the customer directly affects their value.

### 5.4.2 Normalising RFM Scores and Obtaining CLV

The next step was to normalise the R, F, and M scores on a 5-point scale using the following formula (Manoj, 2021):

where minVal and maxVal are the observed maximum values.

This is to prevent a single metric from dominating the overall score rating by ensuring all scores were on the same scale. Finally, we obtained their CLV using the following formula:

### 5.4.3 Segmenting Customers based on CLV Using K-Means Clustering

K-Means clustering is an unsupervised learning clustering method that partitions a dataset into K distinct, non-overlapping clusters (James et al., n.d.). A good k-means clustering indicates that the within-cluster variation is as small as possible. This is measured using the sum of squares where the objective is to partition all observations into clusters such that the total within-cluster sum of squares is as small as possible. The process to minimise the objective starts from randomly assigning a number serving as the initial cluster assignment followed by iterating until the cluster assignment stops changing. The process can be visualised in Appendix F.1. (Hastie, n.d.)

K-Means clustering was used to group our CLV values with the intention of recommending groups of customers for businesses to focus on. The elbow method was used to determine the optimal number of clusters (Refer to Appendix F.2). The optimal numbers for the UCI and Olist datasets were 2 & 3, and 3 & 4 respectively. Thus, we chose to split the CLV into 3 clusters, namely 1 (Low), 2 (Moderate), and 3 (High).

Since K-Means is a multivariate clustering method, we have included the library Ckmeans.1d.dp which follows the K-Means algorithm while optimising it for 1-dimensional variables. A comparison between Ckmeans.1d.dp and K-Means with nstart = 25 was made and we found that Ckmeans.1d.dp had a lower total within-cluster sum of squares for UCI and no difference for Olist (Refer to Appendix F.3). The lower total within-cluster sum of squares was due to the different clustering conducted by both models (Refer to Appendix F.4). Since we prioritise the model with lower total within-cluster sum of squares, we decided to use the results from Ckmeans.1d.dp.

We discovered 3 rows of NA values belonging to cluster 1. We omitted it since cluster 1 has the highest number of rows and omitting will not significantly affect the analysis.

## 5.5. Datasets Used

In predicting CLV, we utilise the two UCI and Olist datasets. For the machine learning models in Sections 5.6, 5.7, and 5.8, we realised that both datasets were not balanced, as Cluster 3 (high-value customers) were the minority class. We hence balanced the datasets through controlled sampling of the middle and majority class.

### 5.5.1 Importance of Balancing to Solving our Business Problem

Owing to the Pareto Principle observed in most forms of business, we expect most real-life datasets to be similarly imbalanced as that observed in the UCI and Olist dataset, where a huge majority of customers are not valuable (Cluster 1) with a fraction that are high value (Cluster 3). (TPP Wholesale, 2016) Since our business objective is to maximise CLV and inform business strategy, we would prefer models with a high accuracy in predicting Cluster 3 customers.

## 5.6 Random Forest

Random Forest for classification was conducted on both datasets. We summarise our findings below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Trainset Confusion Matrix | Trainset Accuracy | Testset Confusion Matrix | Testset Accuracy | Optimal B / RSF\* |
| UCI, Original |  | 99.5%  51.0%  99.9%  **96.8%** |  | 99.5%  51.0%  99.8%  **96.7%** | B=500  RSF=no of vars |
| UCI, Balanced |  | 86.8%  88.8%  100%  **91.9%** |  | 86.8%  88.6%  100%  **87.1%** | B=500  RSF=sqrt(no of vars) |
| Olist, Original |  | 99.4%  30.8%  57.4%  **90.9%** |  | 99.4%  32.0%  54.9%  **91.3%** | B=500  RSF=no of vars |
| Olist, Balanced |  | 58.2%  48.2%  86.5%  **64.3%** |  | 57.9%  52.8%  81.6%  **57.4%** | B=100  RSF=sqrt(no of vars) |

### 5.6.1 Variable Importance

Plots of the random forest variable importance can be found in Appendix E.3.

The UCI dataset shows ‘Country’ as the most important variable by far, being far more important than the second and third most important variables, ‘MonthName’ and ‘DayofWeek’. This difference was more pronounced in the balanced dataset.

A similar trend is seen in the Olist dataset, where the top 5 most important variables were product\_category, freight\_value, review\_score, customer\_state, and payment\_instalments.

### 5.6.2 Accuracies

Models have lower overall Accuracies on balanced datasets. This is due to their notably increased accuracy in predicting customers in Cluster 2 and 3, and corresponding lower accuracy in predicting customers in Cluster 1, which form the majority of our data. Due to reasons stated in [Section 5.5.1](#_heading=h.qyke2pmni8qc), we will prefer the random forest model on the balanced dataset and conclude that balancing the data is important.

### 5.6.3 Hyperparameter Tuning

\*Plots of the error rates against number of trees, as well as a guide on how to interpret them, can be found in Appendix E.4.

We permuted B = (25, 100, 500) with RSF = (1, , ) and the results are tabulated in matrices shown in Appendix E.5. The best RSF value is the (i.e. random forest with no subsetting, only bagging) on the original datasets with no balancing, but on the balanced datasets.

Due to reasons stated in [Section 5.5.1](#_heading=h.qyke2pmni8qc), and the need to decrease correlation of trees (see Appendix E2), we would prefer RSF=*sqrt(no of vars)*. As for B, we will prefer B=500, as that is what 75% of our best-performing models used. This reinforces the idea that R’s default settings of B=500 and RSF=sqrt(no of vars) for classification are the most optimal parameters.

### 5.6.4 Conclusion

Interestingly, the important variables in all four versions of random forest show that the date of purchase is not the most important variable in predicting CLV. There are many other factors such as customer demographics (UCI) and purchasing patterns (Olist) that may influence CLV, offering us evidence that a purely time series-based model is insufficient for CLV prediction.

Our first attempt at a machine learning-based model for CLV prediction was considerably successful in terms of accuracy rate (>80% Cluster 3 prediction on balanced datasets). However, one severe limitation of random forest is the lack of a convenient, in-built way to extract the feature importance of individual categorical predictors. (StackExchange, 2020) This is extremely important in our business problem, as it is not enough to know that ‘Country’ is an important variable in predicting CLV - we require the breakdown of which countries are more important than others as well, so as to inform business strategy. Thus, we posit to utilise logistic regression instead.

## 5.7 Logistic Regression

To find the levels of variable importance, Multinomial Logistic Regression was conducted on both datasets. The models were optimised using backwards elimination and findings will be elaborated in the later sections. The results are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Original Trainset** | | **Balanced Trainset** | |
|  | Confusion Matrix | Accuracy | Confusion Matrix | Accuracy |
| **UCI**  **Train** |  | Cluster 1: 99.61%  Cluster 2: 0%  Cluster 3. 96.78%  Overall: 93.92% |  | Cluster 1: 59.07%  Cluster 2: 63.52%  Cluster 3. 100%  Overall: 74.20% |
| **UCI**  **Test** |  | 1: 99.60%  2: 0%  3: 97.13%  Overall: 93.80% |  | Cluster 1: 57.21%  Cluster 2: 65.34%  Cluster 3: 100%  Overall: 58.52% |
| **Olist**  **Train** |  | Cluster 1: 98.39%  Cluster 2: 40.97%  Cluster 3: 36.83%  Overall: 91.11% |  | Cluster 1: 60.88%  Cluster 2: 45.99%  Cluster 3: 71.76%  Overall: 59.54% |
| **Olist**  **Test** |  | Cluster 1: 98.40%  Cluster 2: 41.58%  Cluster 3: 26.70%  Overall: 91.34% |  | Cluster 1: 53.00%  Cluster 2: 37.56%  Cluster 3: 65.53%  Overall: 51.29% |

#### 5.7.1 Variable Importance

We conducted two methods to determine the variable importance. Firstly, confidence intervals. The confidence intervals for UCI were found to exclude 0, showing statistical significance (Refer to Appendix G.1). Olist also showed similar results as all but one confidence intervals for the variables excluded 0 (Refer to Appendix G.2). We found a NaN for customer\_state, PE level for cluster 3. This shows that the level PE has no significance in predicting cluster 3. The second method to ascertain our findings was to calculate the p-value. Through our calculations, we found that most variables in the UCI trainset had a p-value of 0, or less than 0.05 and most variables in the Olist dataset had a p-value of less than 0.05 (Refer to Appendix G.3 and G.4). Backward elimination was conducted afterwards and there were no changes to the UCI trainset. On the other hand, the Olist trainset only had a variable level removed. This shows that almost all variables in both datasets are significant, resulting in the none-to-small changes through backwards elimination. The same findings were discovered on the balanced trainset, further reinforcing our findings on variable significance.

#### 5.7.2 Accuracies

The Logistic Regression accuracies were lower when the model is trained with the balanced trainset. However, the accuracy in predicting customers who fall under cluster 3 have notably increased. Due to reasons stated in [Section 5.5.1](#_heading=h.qyke2pmni8qc), we will prioritize the model on the balanced dataset.

#### 5.7.2 Conclusion

Multinomial Logistic provided us with important insights into variable importance and allowed us to recommend on which variable level to focus on. However, as most variables were found to be statistically significant with p-values of less than 0 and confidence interval that excludes 0, we were unable to provide definite recommendations on the important variable levels. To further analyse the variable levels, we extended our analysis to include MARS.

## 5.8 Multivariate Adaptive Regression Splines (MARS)

For further analysis of variable importance and comparison to the earlier models, MARS model was built on both datasets. The findings are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Original Trainset** | | **Balanced Trainset** | |
|  | Confusion Matrix | Accuracy | Confusion Matrix | Accuracy |
| **UCI**  **Train** |  | Cluster 1: 99.48%  Cluster 2: 5.06%  Cluster 3. 97.27%  Overall: 94.09% |  | Cluster 1: 66.04%  Cluster 2: 71.40%  Cluster 3. 100%  Overall: 79.15% |
| **UCI**  **Test** |  | Cluster 1: 99.49%  Cluster 2: 5.58%  Cluster 3: 97.47%  Overall: 94.03% |  | Cluster 1: 66.31%  Cluster 2: 71.64%  Cluster 3: 100%  Overall: 67.28% |
| **Olist**  **Train** |  | Cluster 1: 99.56%  Cluster 2: 6.53%  Cluster 3: 3.05%  Overall: 87.79% |  | Cluster 1: 64.69%  Cluster 2: 49.05%  Cluster 3: 81.68%  Overall: 65.14% |
| **Olist**  **Test** |  | Cluster 1: 99.56%  Cluster 2: 6.94%  Cluster 3: 2.43%  Overall: 88.18 % |  | Cluster 1: 57.41%  Cluster 2: 39.29%  Cluster 3: 65.53%  Overall: 55.35% |

#### 5.8.1 Parameter Tuning

We tuned the MARS nfold and degree parameters to determine a higher GRSq and accuracy. We prioritized a higher overall GRSq to prevent overfitting of our model, followed by accuracy in our predictions. After tuning these parameters, we found that a degree of 2 provides a higher GRSq for our models on UCI and Olist trainsets, as well as each of the balanced trainsets (Refer from Appendix H.1 to H.8). We also found that a degree of 2 provides our models with better accuracy too. The nfold parameter did not affect the GRSq nor accuracy, thus we decided to use the default settings for nfold.

#### 5.8.2 Variable Importance

We were able to extract the variable importance using the evimp function (Refer from Appendix H.9 to H.12). The function allowed us to determine the importance of the variable and its level by measuring its GCV and RSS values. Based on the results, we found that the top 3 important variables of each dataset (#1 on left, #3 on right) are:

|  |  |
| --- | --- |
| **Dataset** | **Top 3 Variable Name/Levels** |
| UCI Original Trainset | |  |  |  | | --- | --- | --- | | InvoiceDate\_DayofMonth | InvoiceDate\_MonthName, August | Country, United Kingdom | |
| UCI Balanced Trainset | |  |  |  | | --- | --- | --- | | InvoiceDate\_MonthName, September | InvoiceDate\_DayofMonth | Country, Finland | |
| Olist Original Trainset | |  |  |  | | --- | --- | --- | | payment\_type, voucher | product\_category\_name\_english, housewares | product\_category\_name\_english, construction\_tools\_lights | |
| Olist Balanced Trainset | |  |  |  | | --- | --- | --- | | product\_category\_name\_english, watches\_gifts | product\_category\_name\_english, home\_construction | product\_category\_name\_english, baby | |

#### 5.8.3 Accuracies

Similar to the previous models, MARS also showed a reduction in overall accuracy while having a higher accuracy on cluster 3 predictions on the balanced trainset. However, we found that MARS surpassed the earlier models on the UCI balanced trainset as it predicted cluster 3 with 100% accuracy. This provides an answer for our business problem specified in [Section 5.5.1](#_heading=h.qyke2pmni8qc) where we aim to prioritise high CLV value customers.

#### 5.8.4 Conclusion

MARS was able to provide deeper insights into our business solution by breaking down the variable importance to its levels. By using MARS, we will be able to specify which product category that businesses should focus on to maximise their CLV. MARS also provided the highest accuracy in predicting cluster 3 customers on the UCI dataset. This provides valuable insights as the model can successfully predict a cluster 3 customer for businesses while recommending how to target them. Since MARS has met our business needs, we will be prioritising MARS as our recommended model.

## 5.9 Optimising Customer Lifetime Value Through Machine Learning

We developed models on two different datasets as we would like to compare the performance of models on a DTC dataset versus an ecommerce platform dataset.

In the time-series model, we can interpret the dataset that captures a longer period as belonging to that of the e-commerce platform, since platform marketplaces have been around longer than DTC channels. This refers to the Olist (and Kaggle) dataset, and their general better prediction accuracy than that for UCI, which captures data over a shorter period.

Beyond looking at Customer Lifetime Value from a temporal perspective, we also looked at it from a spatial perspective, assigning CLV by customer instead of by time period and grouping customers into 3 distinct value clusters. From this temporal perspective, we will interpret instead the UCI dataset to belong to the e-commerce platform, since the UCI dataset has more rows, capturing a larger volume of transactions.

A DTC company can follow and match its variables to an e-commerce platform to gain similar insights. However, a DTC company will not be able to match the consumer purchase-traffic on an e-commerce platform due to the difference in variety on both platforms. Therefore, an e-commerce platform’s dataset will always provide more data points or rows for analysis.

Referring to [Section 1.3.3](#_heading=h.up0oy2e57vao), the models have provided sufficient insights for business to optimise their CLV and increase their margins. Our models have shown better performance on an ecommerce-type dataset as measured by the accuracies of the model on the UCI dataset against the Olist dataset. The Random Forest and Logistic Regression models have both shown high accuracy in predicting cluster 3 customers, but lack foresight when it comes to variable specifics. Here, MARS has shown superiority over the two models with its 100% cluster 3 accuracy while being able to view down to the specifics as to which variable level is important. This is essential and significant for businesses as direct and effective recommendations can be made to improve their margins and optimise their CLV (e.g. focus on customers from a specific country). Since our purpose is to assist DTC companies as an ecommerce platform, we recommend the MARS model due to its 100% cluster 3 accuracy on the balanced trainset while being able to show detailed information on variable importance which can greatly impact a business.

# 6. Conclusion

Our findings conclude that machine learning and analytics are successful in solving our business problems and meeting the business opportunity of ecommerce platforms providing more benefits than DTC companies. Our justifications are as follows:

Association Rules was the first technique introduced to meet our business opportunity. The technique was used to derive associations across different products from the same seller. This gives sellers the information on their product complements that allows them to group their product together for higher overall sales. The technique also provided associative insights into products across different sellers which is effective in increasing marginal sales for individual sellers. An ecommerce platform offers variety across different products, sellers or companies that cannot be matched by a DTC company. This shows that an ecommerce platform is superior and can provide better insights into customer purchase patterns than a DTC company.

Our recommended models for the second business opportunity are Quantile Regression and MARS. Quantile Regression showed a higher accuracy at the 50th percentile which provides significant insights into high customer ratings. We were able to determine the factors that resulted in high customer ratings and this will be constructive for businesses in maintaining high ratings. However, since the percentile cannot be specified at the time of prediction, MARS, which showed a higher overall accuracy, was introduced as the second recommended model for customers to more accurately predict their customer ratings. Overall, MARS and Quantile Regression work hand-in-hand in predicting the customer rating for sellers on ecommerce platforms and determining the significant variables for high ratings. This can be better established in ecommerce platforms as ecommerce platforms have a larger customer base, indicating that they can receive more constructive ratings that correlates to the significant variables. This shows that an ecommerce platform has a better advantage over DTC companies due to the depth of insights it can gain from a larger customer base.

Our third business opportunity focused on optimising clv using various models and concluded MARS to be the best model. Here, we prioritised accuracy in predicting high clv customers since they will be the target for all businesses. As an ecommerce platform, such insights will be beneficial to sellers or companies in determining the group of customers to target. The MARS model successfully fulfills this need as it was able to predict cluster 3 customers with 100% accuracy. Furthermore, MARS was able to shed light on variable importance which can give businesses more information on their next steps to optimising their clv. MARS was also found to perform better on an ecommerce dataset with more data points. This shows that ecommerce platforms have an advantage over DTC companies in the use of machine learning as the MARS model was able to perfectly predict high clv customers on an ecommerce dataset while providing insights into variable importance.

In conclusion, we find that ecommerce platforms have an explicit advantage over DTC companies due to variety, depth and accuracy.

# References

Baesens, B. (2017, April 10). *What is the lift value in association rule mining?* DataMiningApps. Retrieved April 3, 2022, from https://www.dataminingapps.com/2017/04/what-is-the-lift-value-in-association-rule-mining/

Big Commerce. (2022). *Why the Digitally Native Vertical Brand is Claiming Control*. BigCommerce. Retrieved April 3, 2022, from https://www.bigcommerce.com/articles/direct-to-consumer/dnvbs/#are-dnvbs-the-future-of-retail-

D, A. (2018, September 14). *Association Rule*. GeeksforGeeks. Retrieved April 3, 2022, from https://www.geeksforgeeks.org/association-rule/

DC Velocity. (2022, January 17). *SEKO Ecommerce launch connects primary global supply chain to ecommerce*. DC Velocity. Retrieved April 3, 2022, from https://www.dcvelocity.com/articles/53486-seko-ecommerce-launch-connects-primary-global-supply-chain-to-ecommerce

Dobilas, S. (2020, November 28). *MARS: Multivariate Adaptive Regression Splines — How to Improve on Linear Regression?* Towards Data Science. Retrieved April 3, 2022, from https://medium.com/towards-data-science/mars-multivariate-adaptive-regression-splines-how-to-improve-on-linear-regression-e1e7a63c5eae

Dobilas, S. (2020, November 28). *MARS: Multivariate Adaptive Regression Splines — How to Improve on Linear Regression?* Towards Data Science. Retrieved April 3, 2022, from https://medium.com/towards-data-science/mars-multivariate-adaptive-regression-splines-how-to-improve-on-linear-regression-e1e7a63c5eae

Dye, S. (2020, February 12). *Quantile Regression. When performing regression analysis, It… | by Steven Dye*. Towards Data Science. Retrieved April 3, 2022, from https://towardsdatascience.com/quantile-regression-ff2343c4a03

eDesk. (2021, November 9). *DTC eCommerce: Why Direct To Consumer Is The Future*. eDesk. Retrieved April 3, 2022, from https://www.edesk.com/blog/dtc-ecommerce-direct-to-consumer/

Geldman, A. (2021, June 10). *Online Marketplaces in the USA: Amazon is Not The Only Show in Town*. Web Retailer. Retrieved April 3, 2022, from https://www.webretailer.com/b/online-marketplaces-usa/

Goulart, K. (2020). *What is omnichannel? - Definition from WhatIs.com*. TechTarget. Retrieved April 3, 2022, from https://searchcustomerexperience.techtarget.com/definition/omnichannel

Hastie, T. (n.d.). *An Introduction to Statistical Learning*. Trevor Hastie. Retrieved April 3, 2022, from https://hastie.su.domains/ISLR2/ISLRv2\_website.pdf

Hastie, T., & Tibshirani, R. (n.d.). *Notes on the earth package*. Stephen Milborrow Homepage. Retrieved April 3, 2022, from http://www.milbo.org/doc/earth-notes.pdf

Hastie, T., Tibshirani, R., James, G., & Witten, D. (2013). *An Introduction to Statistical Learning: With Applications in R* (G. James, Ed.). Springer New York.

Hyndman, R. (2016, July 29). *R: Why does it mean when the MASE of a forecast model is NaN?* StackOverflow. https://stackoverflow.com/questions/38649068/r-why-does-it-mean-when-the-mase-of-a-forecast-model-is-nan

Hyndman, R. J. (n.d.). *ANOTHER LOOK AT FORECAST-ACCURACY METRICS FOR INTERMITTENT DEMAND*. Rob J Hyndman. Retrieved April 3, 2022, from https://robjhyndman.com/papers/foresight.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). *An Introduction to Statistical Learning*. Trevor Hastie. Retrieved April 3, 2022, from https://hastie.su.domains/ISLR2/ISLRv2\_website.pdf

Kassambara. (2018, November 3). *Linear Regression Assumptions and Diagnostics in R: Essentials - Articles*. STHDA. Retrieved April 3, 2022, from http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

Koss, H. (2022). *27 Direct To Consumer (DTC) Brands To Know*. Built In. Retrieved April 3, 2022, from https://builtin.com/marketing/direct-to-consumer-brands

Lebow, S. (2021, August 19). *Worldwide ecommerce continues double-digit growth following pandemic push to online*. eMarketer. Retrieved April 3, 2022, from https://www.emarketer.com/content/worldwide-ecommerce-continues-double-digit-growth-following-pandemic-push-online

Manoj. (2021, April 27). *RFM Analysis | Cutomer Lifetime Value using RFM Analysis*. Analytics Vidhya. Retrieved April 3, 2022, from https://www.analyticsvidhya.com/blog/2021/04/customer-lifetime-value-using-rfm-analysis/

McKinsey & Co. (2020, October 5). *COVID-19 digital transformation & technology*. McKinsey. Retrieved April 3, 2022, from https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever

Milborrow, S. (2020, Jul 29). *Predicting Probabilities of non-binary outcomes with multivariate adaptive regression splines (MARS) in R*. Stats StackExchange. https://stats.stackexchange.com/questions/479480/predicting-probabilities-of-non-binary-outcomes-with-multivariate-adaptive-regre

Nithyakumar, K. (2022, March 10). *What is Customer Lifetime Value (CLV) | Definition, Formula & Calculation*. Zoho. Retrieved April 3, 2022, from https://www.zoho.com/subscriptions/guides/what-is-customer-lifetime-value-clv.html

qualtrics. (2022). *What is Customer Lifetime Value (CLV) ?* Qualtrics. Retrieved April 3, 2022, from https://www.qualtrics.com/uk/experience-management/customer/customer-lifetime-value/

Salpini, C. (2021, March 23). *How Nike is using DTC and data to expand its empire*. Retail Dive. Retrieved April 3, 2022, from https://www.retaildive.com/news/how-nike-is-using-dtc-and-data-to-expand-its-empire/596602/

Scott, G. (n.d.). *Time Series Definition*. Investopedia. Retrieved April 3, 2022, from https://www.investopedia.com/terms/t/timeseries.asp

Segal, T., & Kvilhaug, S. (n.d.). *Recency, Frequency, Monetary Value (RFM) Definition*. Investopedia. Retrieved April 3, 2022, from https://www.investopedia.com/terms/r/rfm-recency-frequency-monetary-value.asp

Shruthi. (2021, August 24). *“The Rise of Digitally Native Brands (DNVB)”*. DataWeave. Retrieved April 3, 2022, from https://dataweave.com/blog/dnvb-direct-to-consumer

StackExchange. (2020, April 2). *R: Importance of Categorical Variables in Random Forests*. Cross Validated. Retrieved April 3, 2022, from https://stats.stackexchange.com/questions/457953/r-importance-of-categorical-variables-in-random-forests

Straight, B. (2021, January 13). *Global e-commerce sales jump 24% in December*. FreightWaves. Retrieved April 3, 2022, from https://www.freightwaves.com/news/global-e-commerce-sales-jump-24-in-december

Sularia, S. (2021, January 22). *How Shopify Is Shifting The E-Commerce Landscape*. Forbes. Retrieved April 3, 2022, from https://www.forbes.com/sites/forbestechcouncil/2021/01/22/how-shopify-is-shifting-the-e-commerce-landscape/

Thomas, S. (2019, December 11). *Data Cleaning in Machine Learning: Best Practices and Methods*. eInfochips. Retrieved April 3, 2022, from https://www.einfochips.com/blog/data-cleaning-in-machine-learning-best-practices-and-methods/

TIBCO. (n.d.). *Multivariate Adaptive Regression Splines (MARSplines) Overview*. TIBCO Product Documentation. Retrieved April 3, 2022, from https://docs.tibco.com/data-science/GUID-ED6E1533-EAE7-4C5B-ACFD-DCC20D894C44.html

Tin, K. H. (n.d.). *Random Decision Forests*. Retrieved April 3, 2022, from https://web.archive.org/web/20160417030218/http://ect.bell-labs.com/who/tkh/publications/papers/odt.pdf

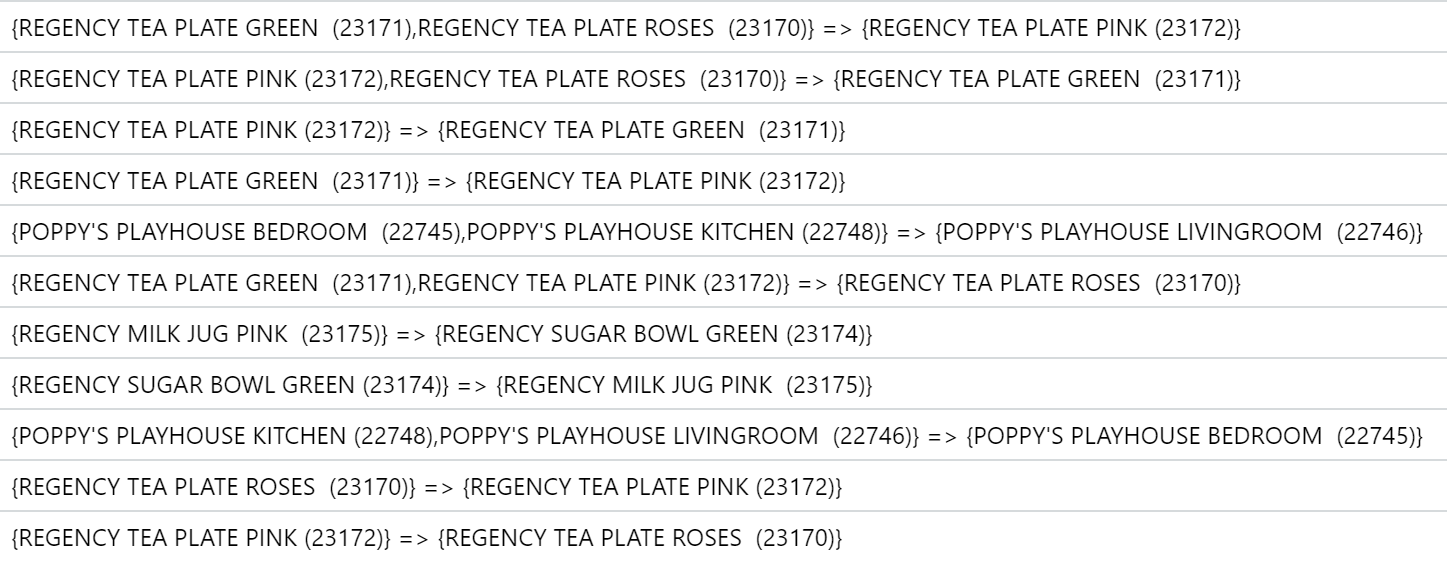
TPP Wholesale. (2016, August 29). *Why customer lifetime value matters*. TPP Wholesale. Retrieved April 3, 2022, from https://www.tppwholesale.com.au/blog/why-customer-lifetime-value-matters/

Verdon, J. (2021, April 27). *Global E-Commerce Sales To Hit $4.2 Trillion As Online Surge Continues, Adobe Reports*. Forbes. Retrieved April 3, 2022, from https://www.forbes.com/sites/joanverdon/2021/04/27/global-ecommerce-sales-to-hit-42-trillion-as-online-surge-continues-adobe-reports/

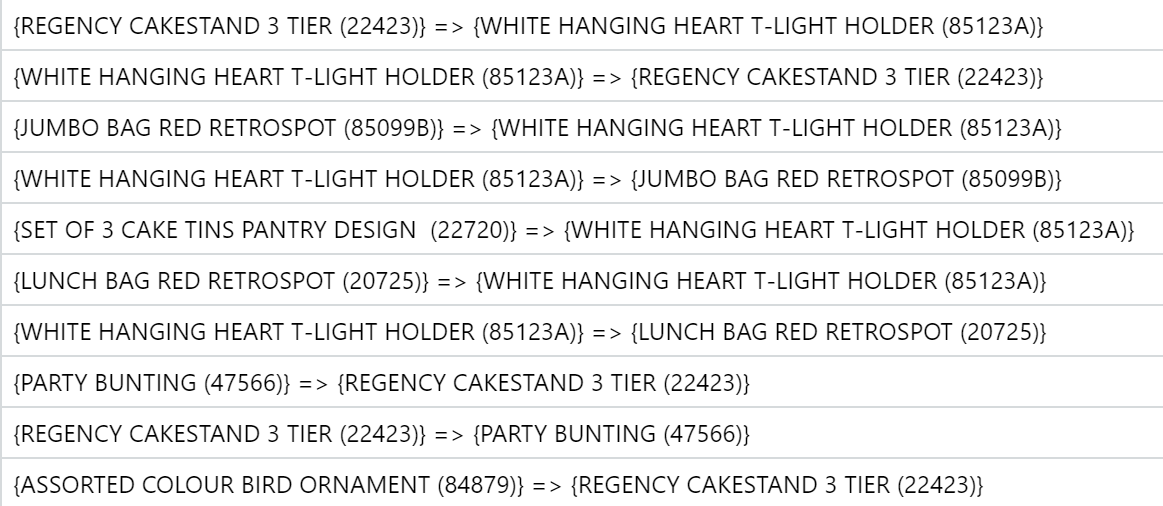
Walk, T. (2021, June 17). *What can DTCs gain from online marketplaces?* Retail Dive. Retrieved April 3, 2022, from https://www.retaildive.com/news/what-can-dtcs-gain-from-online-marketplaces/601097/

# Appendix A: Association Rules

### A.1 Top 10 Association Rules in Terms of Lift Value (UCI Dataset)

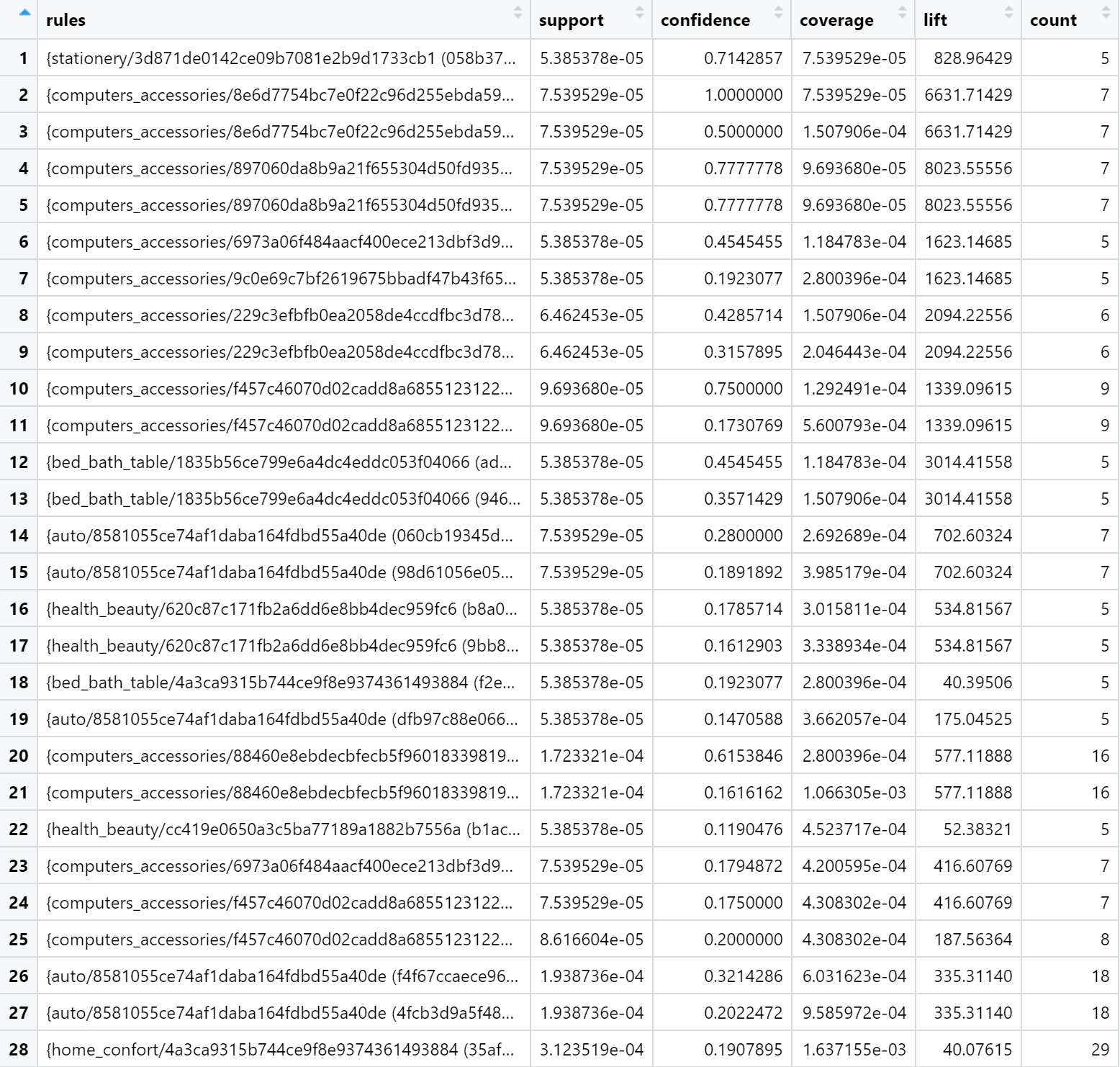


### A.2 Bottom 10 Association Rules in Terms of Lift Value (UCI Dataset)



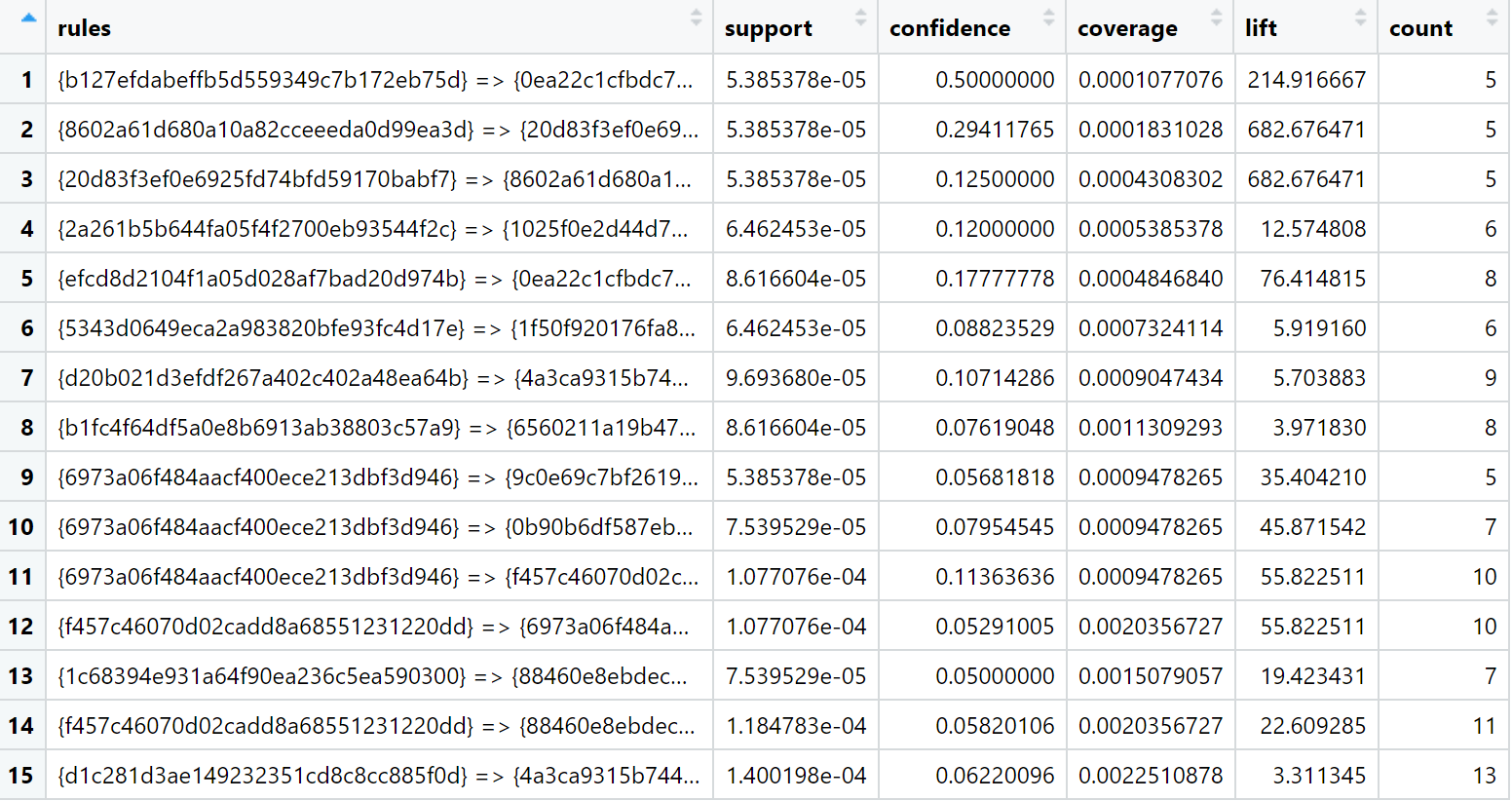
### 

### A.3 Association Rules between Products (Olist Dataset)



### 

### A.4 Association Rules between Sellers (Olist Dataset)

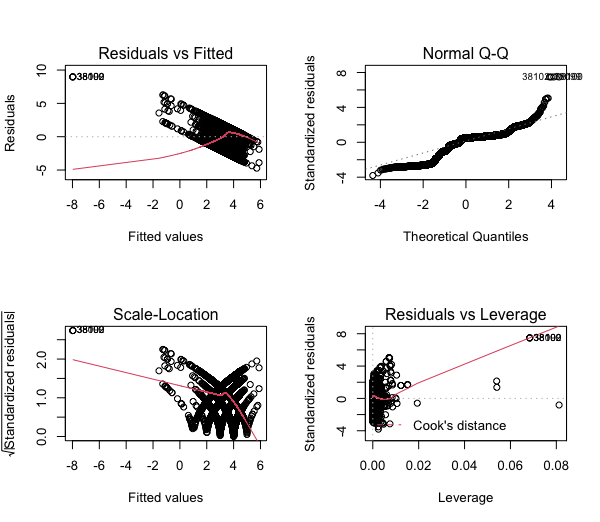


# Appendix B: Linear Regression

### B.1 Derived Variables

|  |  |
| --- | --- |
| del\_time (days) | order\_delivered\_customer\_date - order\_approved\_at |
| est\_del\_time (days) | order\_estimated\_delivery\_date - order\_approved\_at |
| delta\_time (days) | order\_estimated\_delivery\_date - order\_delivered\_customer\_date |
| Late (boolean) | if delta\_time < 0, Late = 1, else = 0 |
| total\_price ($) | price + freight\_value |
| freight\_ratio (<= 1) | freight\_value / price |
| purchase\_day\_of\_week (day name) | day where order is approved at (order\_approved\_at) |

### B.2 Diagnostic Plots

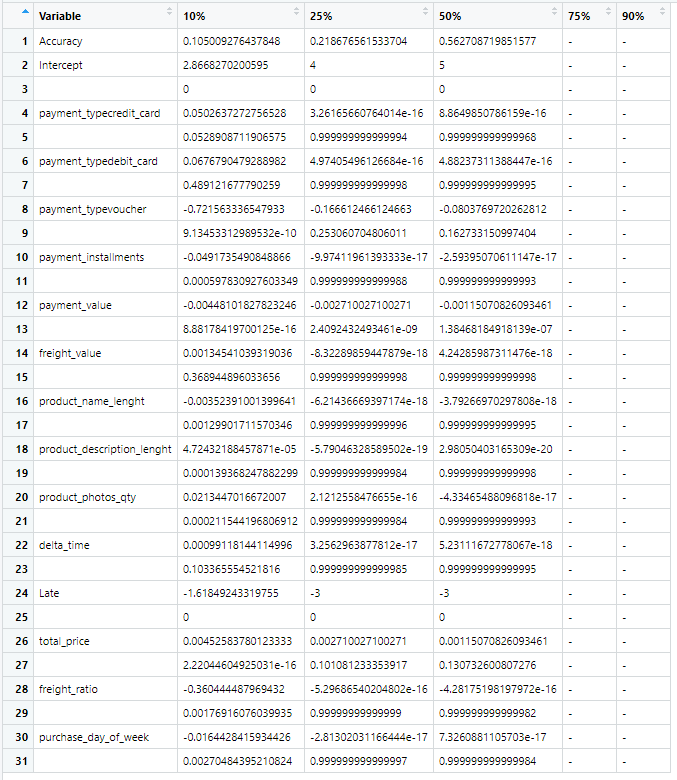


# Appendix C: Quantile Regression

### C.1 Quantile Regression Result on the Original Variables

### 

### C.2 Quantile Regression Result with Additional Derivative Variables



# Appendix D: Time Series

### D.1 Exponential Smoothing Methods

There are three broad methods of time-series forecasting: Moving Average Method, Exponential Smoothing Methods, and Autocorrelation Methods such as Autoregressive Integrated Moving Average, or ARIMA. In this paper, we only explore Exponential Smoothing Methods.

As mentioned, we explore three different time series models. These three models are all forms of Exponential Smoothing:

|  |  |
| --- | --- |
| SES: Simple Exponential Smoothing | Without trend and seasonality |
| Holt’s Method | With trend but no seasonality |
| Holt-Winters’ Method | With seasonality and possibly trend |

Additionally, each model is an extension of the model above, e.g. Holt-Winters’ Method is built on Holt’s Method, which is itself built on Simple Exponential Smoothing. Because of this, we are able to present all three models as a single set of mathematical equations, colour-coded by applicability:

|  |  |
| --- | --- |
| a | smoothing constant between 0 and 1 that smooths historical values. |
| Lt | Level of series at time t = estimate of where the deseasonlised series would be at time t if there is no random error |
| Yt | Data value at time t |
| Ft+k | Forecast is the current level + trend, multiplied by the seasonal component. |
| Lt-1 | Captures the entire history of Y to account for the trend component. |
| B | smoothing constant for trend.  If =0, we get Holt’s Method. |
| Tt | Trend at time T, uses level L hence is already deseasonalised. |
| y | smoothing constant for seasonal effect.  If =0, we get Holt’s Method. |
| -M | ensures we subtract the correct amount of time to get to the seasonal effect in the previous time frame, by not making the assumption that the seasonal effect follows the same frequency as time.  e.g. trend is every quarter M=4  e.g. trend is every month M=12 |
| St | Seasonal component S. Remove the Lt to give us the most recent estimate for S, then smooth it against historical data for seasonal component. |

### D.2 Autoregressive Integrated Moving Average (ARIMA) Method

The ARMA method is based on the concept of autocorrelation, where lagged data points can be used to predict the next data point. An ‘autocorrelation by lag k’ refers to correlation between an observation at time t and at time t-k.

As mentioned, there are three parameters in ARIMA forecasting: p, d, and q. These correspond to the autoregressive, integrated, and moving average components of the model respectively.

Autoregressive: parameter ‘p’

By plotting the partial autocorrelation plots for each of the time series models and counting the number of values that exceed the blue dotted threshold line, we can obtain the value for p.

|  |  |
| --- | --- |
|  | UCI dataset.  No values exceed the threshold line.  p = 0 |
|  | Olist dataset.  No values exceed the threshold line.  p = 0 |
|  | Kaggle dataset.  No values exceed the threshold line.  p = 0 |

Integrated: parameter ‘d’

By plotting the original time series plot as well as the diferencing plots for each of the time series models and determining if the plots have stationary mean and variance, we can obtain the value for d.

|  |  |
| --- | --- |
|  | UCI dataset, difference = 0  ‘Width’ of the plot, i.e. variance, seems to be constant throughout the graph.  d = 0 |
|  | Olist dataset, difference = 0  ‘Width’ of the plot, i.e. variance, seems to be constant throughout the graph.  d = 0 |
|  | Kaggle dataset.  Slightly larger variance at the beginning of 2015, where the ‘width’ of the plot is larger than the rest  However, differencing did not seem to resolve this problem - the problem still persisted for d=1 to d=5 (only up to d=3 is shown)  Hence, we will just take d=0. |

Moving Average: parameter ‘q’

By plotting the autocorrelation plots for each of the time series models and counting the number of values that exceed the blue dotted threshold line, we can obtain the value for q.

|  |  |
| --- | --- |
|  | UCI dataset.  1 value exceeds the threshold line.  q = 1 |
|  | Olist dataset.  1 value exceeds the threshold line.  q = 1 |
|  | Kaggle dataset.  1 value exceeds the threshold line.  q = 1 |

### D.3 Plots of Actual and Forecasted Data

|  |  |
| --- | --- |
| SES (UCI) |  |
|
| Holt’s (UCI) |  |
|
| ARIMA (UCI)  (0,0,1) |  |
|
| Auto ARIMA (UCI)  (0,0,0) |  |
|
| SES (Olist) |  |
|
| Holt’s (Olist) |  |
|
| ARIMA (Olist)  Auto ARIMA (Olist)  (0,0,1) for both |  |
|
| Holt’s (Kaggle) |  |
|
| Winter’s (Kaggle) |  |
|
| ARIMA (Kaggle)  (1,0,1) |  |
|
| Auto ARIMA (Kaggle)  (1,0,0, 1,0,0) |  |
|

# Appendix E: Random Forest

### E.1 Bootstrap Aggregating (“bagging”)

Bootstrapping is a general statistical technique that allows one to make inferences about an unknown population with a known sample. It is achieved by repeatedly creating new datasets of the same number of elements n as the original dataset via sampling with replacement.

Bootstrap aggregating, or “bagging”, is then the act of using each bootstrap sample to create one model. The number of bootstrap samples B equals the number of trainsets equals the number of models. We

1. Get B predictions from B models, and finally
2. Aggregate the results to produce the final prediction.

Number of bootstraps/models B is chosen to be a large number. In R, the default is 500 as of writing. We proved in the report that this default is more than sufficient to produce an optimal random forest model.

### E.2 Random Subset Feature selection (“RSF”)

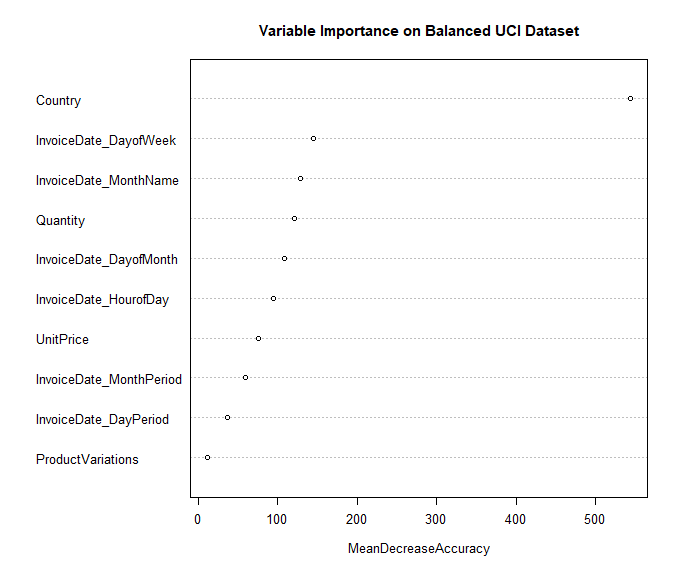
Bagging is not useful in the presence of stable procedures or dominant Xs.

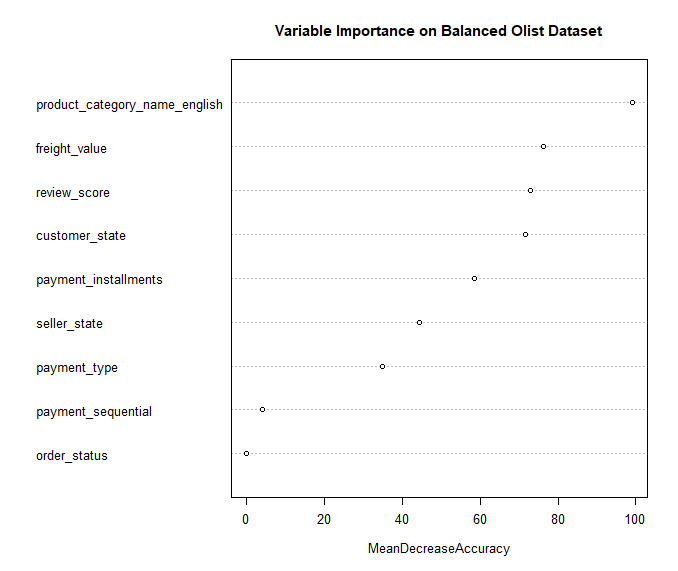
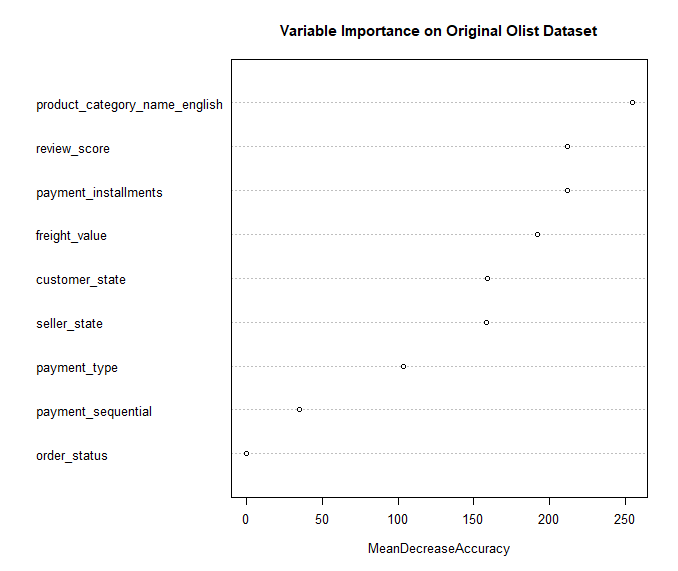
A stable procedure is one where the model generated will be the same on every iteration. When there are dominant X variables, procedures will become more stable since most models will select the dominant X as an important variable. This will result in high correlation of trees.

To decrease correlations of trees, we infuse controlled instability through Random Subset Feature selection, where a different subset of X variables are randomly made available at each split in the decision trees. The non-chosen X variables have no chance to be selected as the best split, and this prevents any dominant Xs from resulting in the same model produced among all bootstrapped trees.

### E.3 Random Forest Variable Importance Plots

‘MeanDecreaseAccuracy’ refers to how much the Random Forest prediction accuracy would decrease if that variable’s data values are permuted, keeping other X variables unchanged.

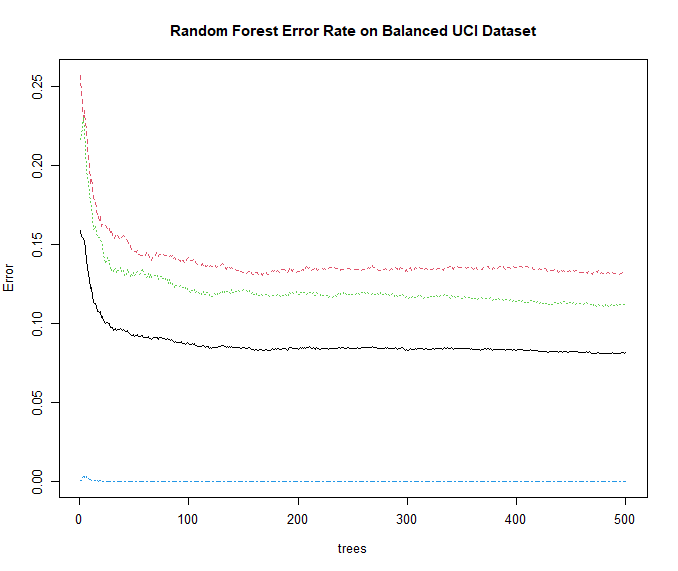
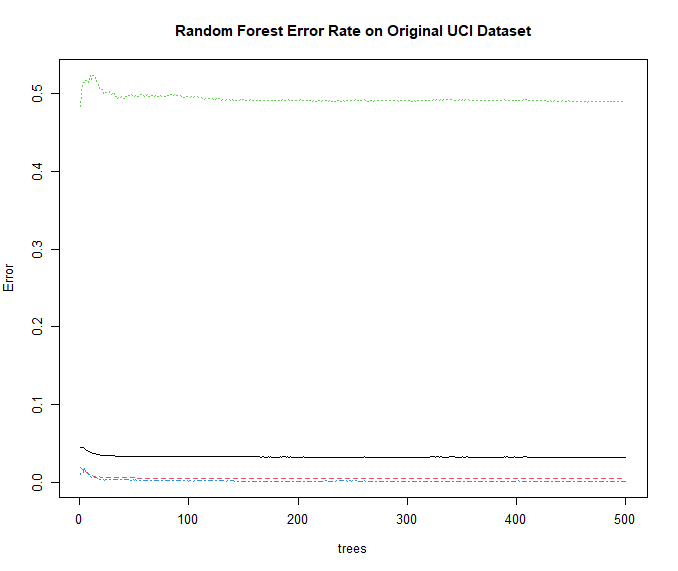


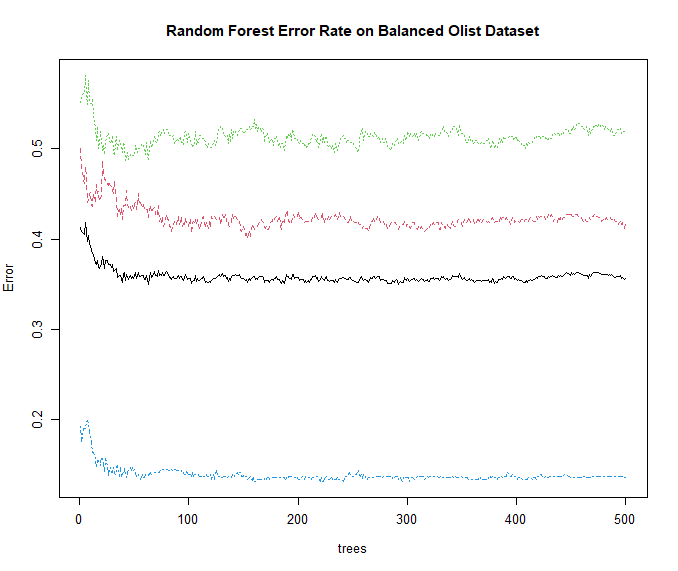
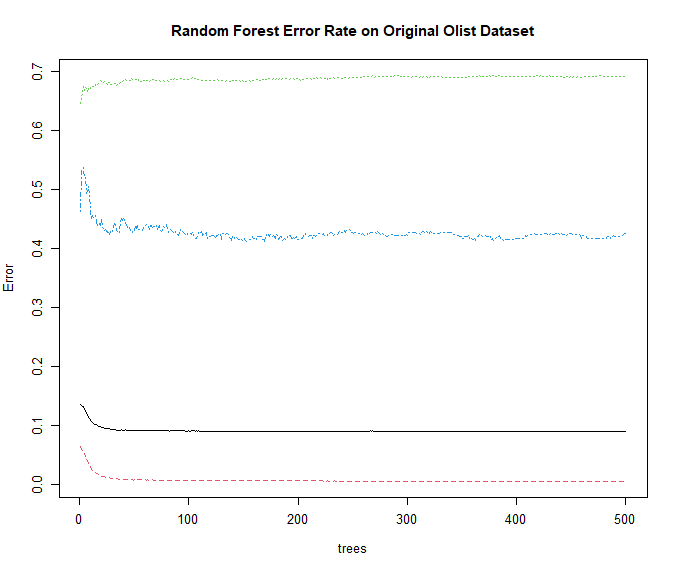
The greater the decrease in accuracy, the more important that variable in our predictions.

### E.4 Random Forest Error Rates against Number of Trees Plots

|  |  |
| --- | --- |
| Green | Error Rate for Cluster 2 |
| Red | Error Rate for Cluster 1 |
| Blue | Error Rate for Cluster 3 |

As expected, as the number of trees increase, the error rates will decrease, up till a plateau. These plots are important, as we must ensure the number of trees we choose must be a sufficiently high number for the error rates to stabilise (become a straight line).





In general, most models have difficulty predicting Cluster 2, as seen by the green line consistently being the highest line (apart from the Balanced UCI dataset case, top right).

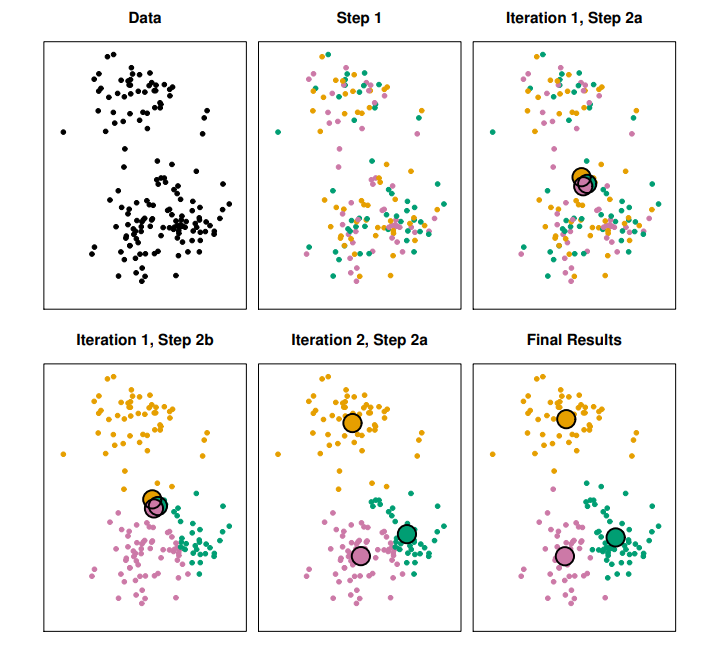
### E.5 Random Forest Hyperparameter Tuning

These show the error rates for random forests built with different hyperparameters. The X axis shows the different number of trees / B / ntry values, and the Y axis shows the different number of variables selected / Random Subset Features selected / RSF values. The lower the number, the better.

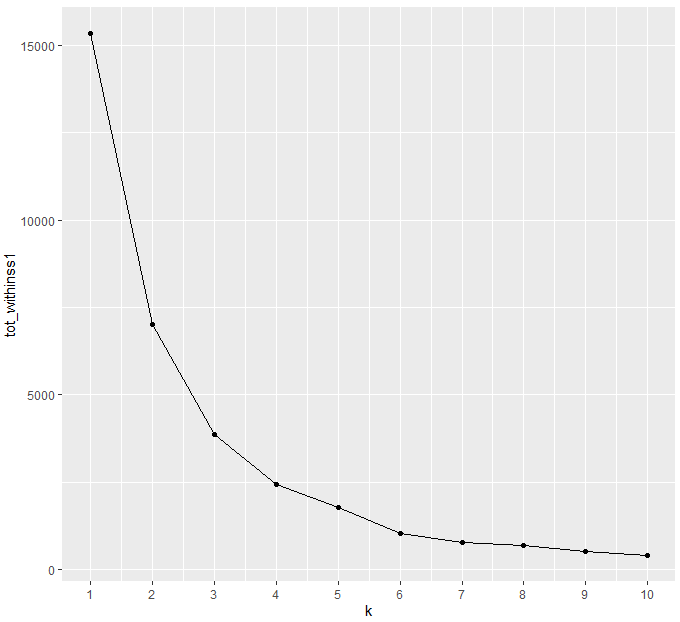
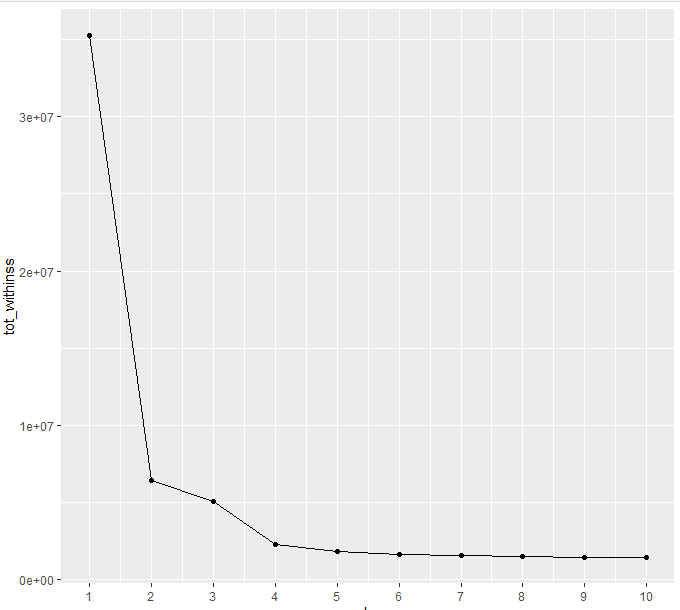
|  |  |
| --- | --- |
| UCI, Original | UCI, Balanced |
| Olist, Original | Olist, Balanced |

# Appendix F: K-Means

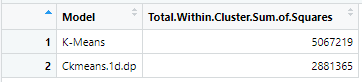
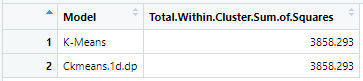
### F.1 K-Means Clustering Process



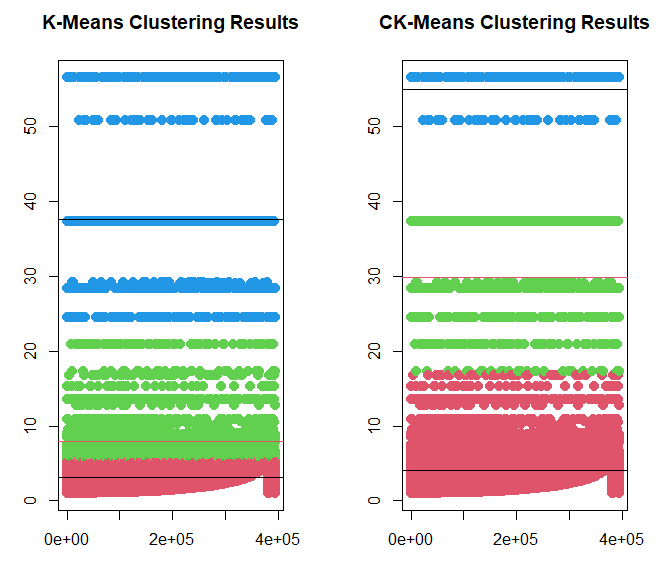
**F.2 K-Means Elbow Method for UCI (Left) and Olist (Right)**



**F.3 K-Means Total-Within Cluster Sum of Squares for UCI (Left) and Olist (Right)**

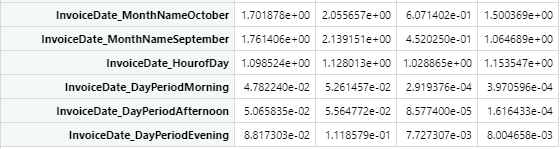
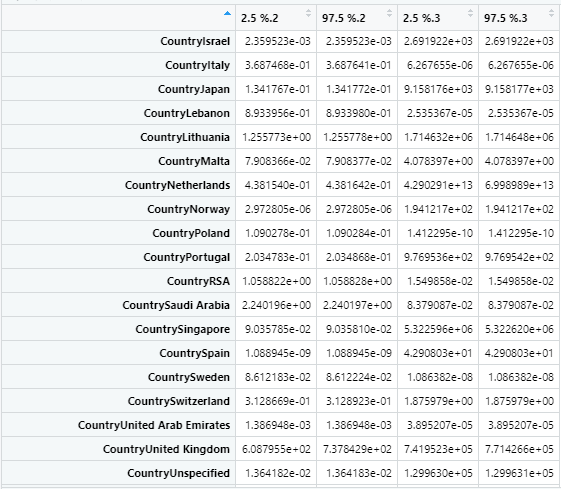
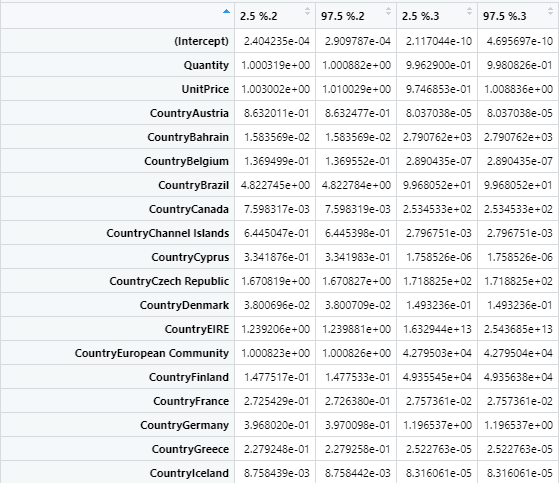
****

**F.4 K-Means Clustering Results for UCI**

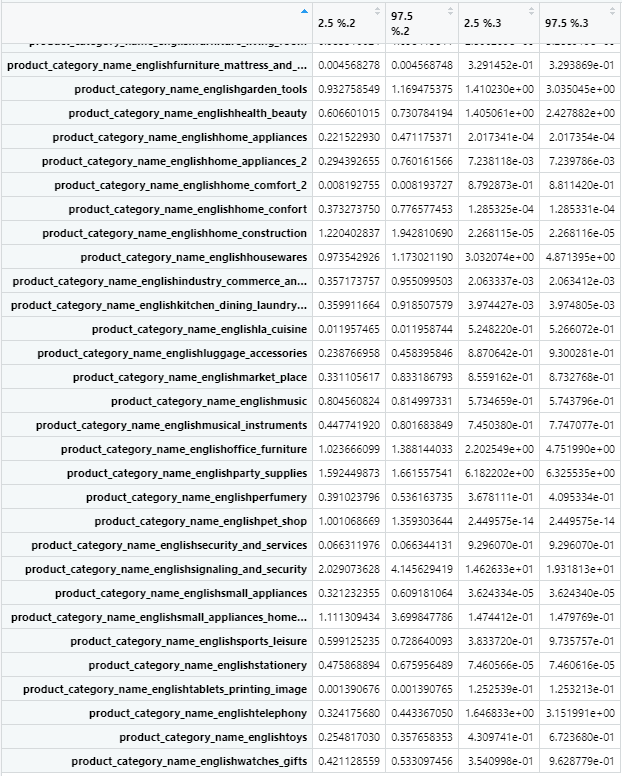
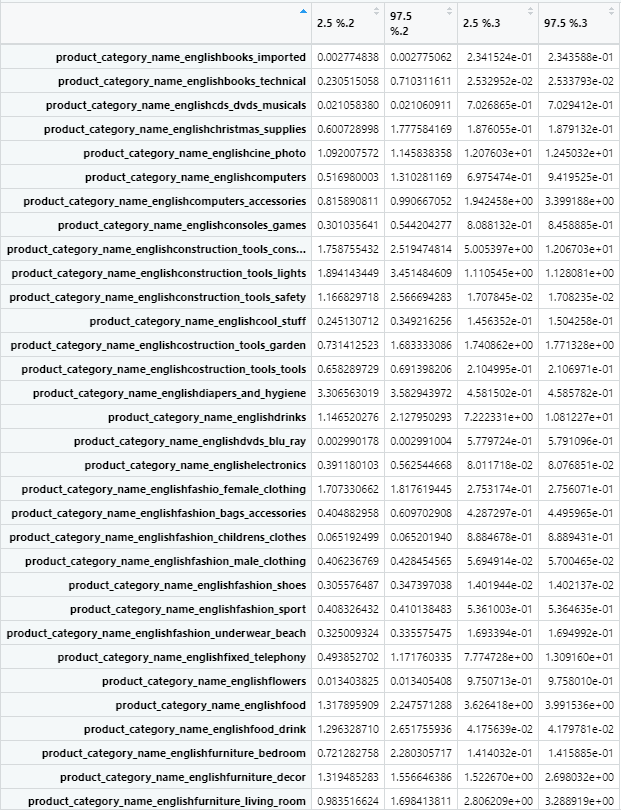
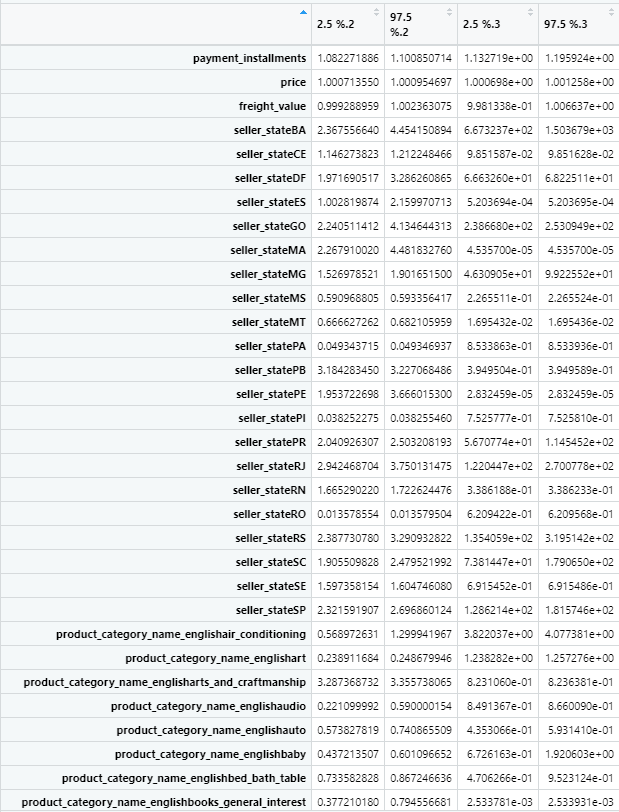
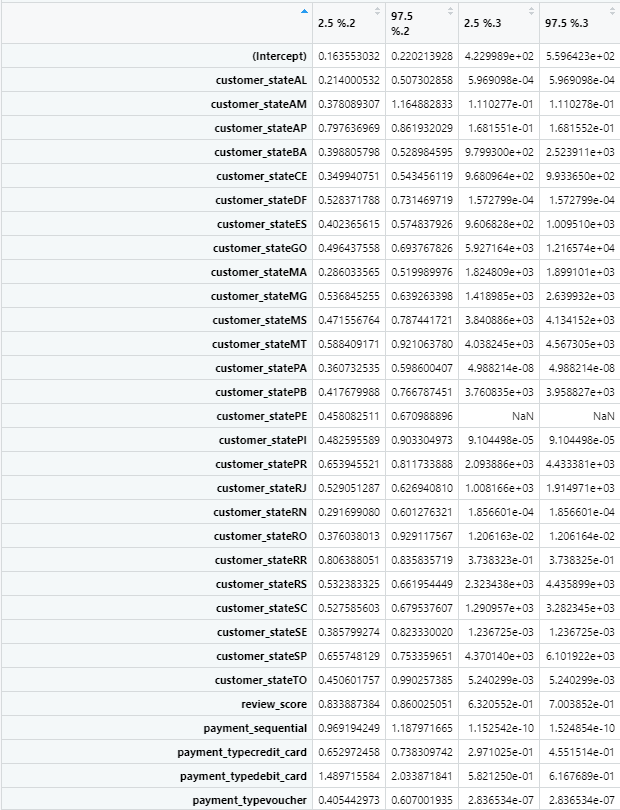


# Appendix G: Logistic Regression

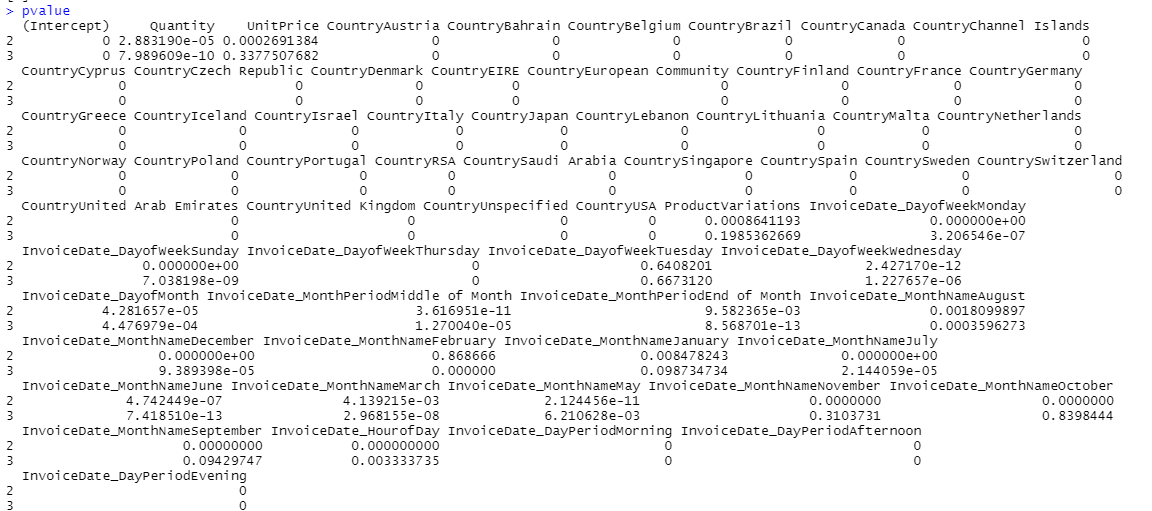
### G.1 Multinomial Confidence Intervals for UCI Original Trainset



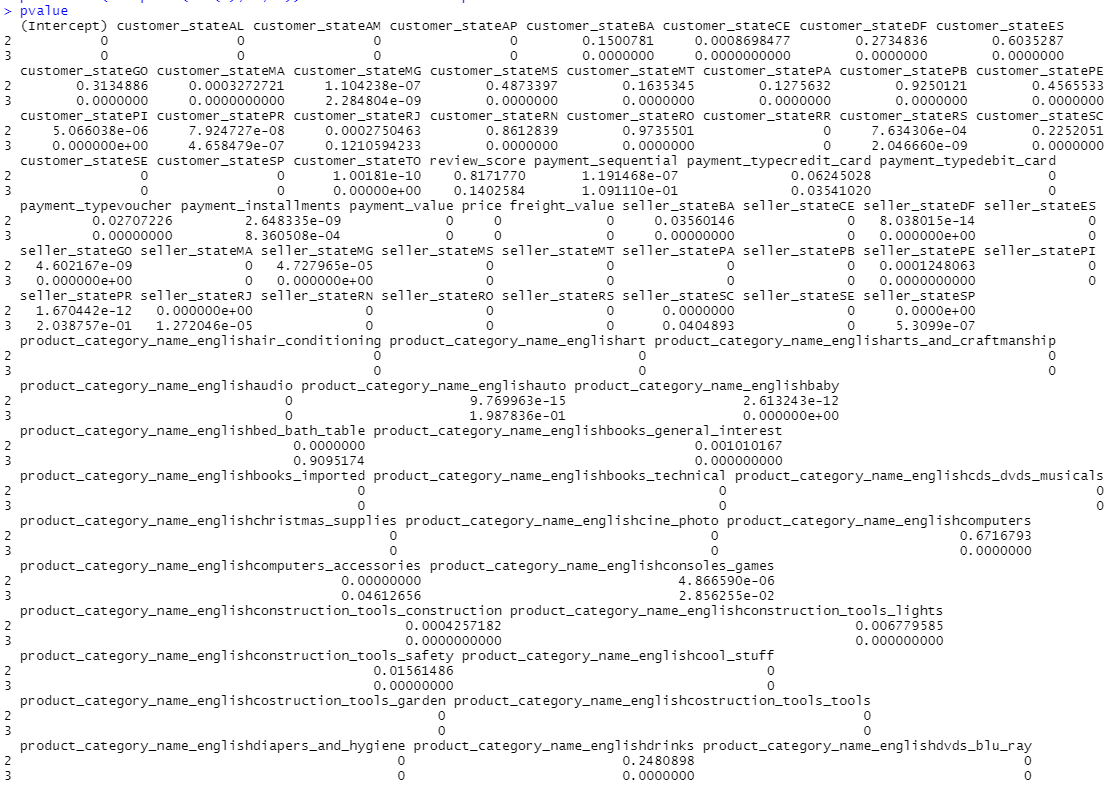
### G.2 Multinomial Confidence Intervals for Olist Original Trainset

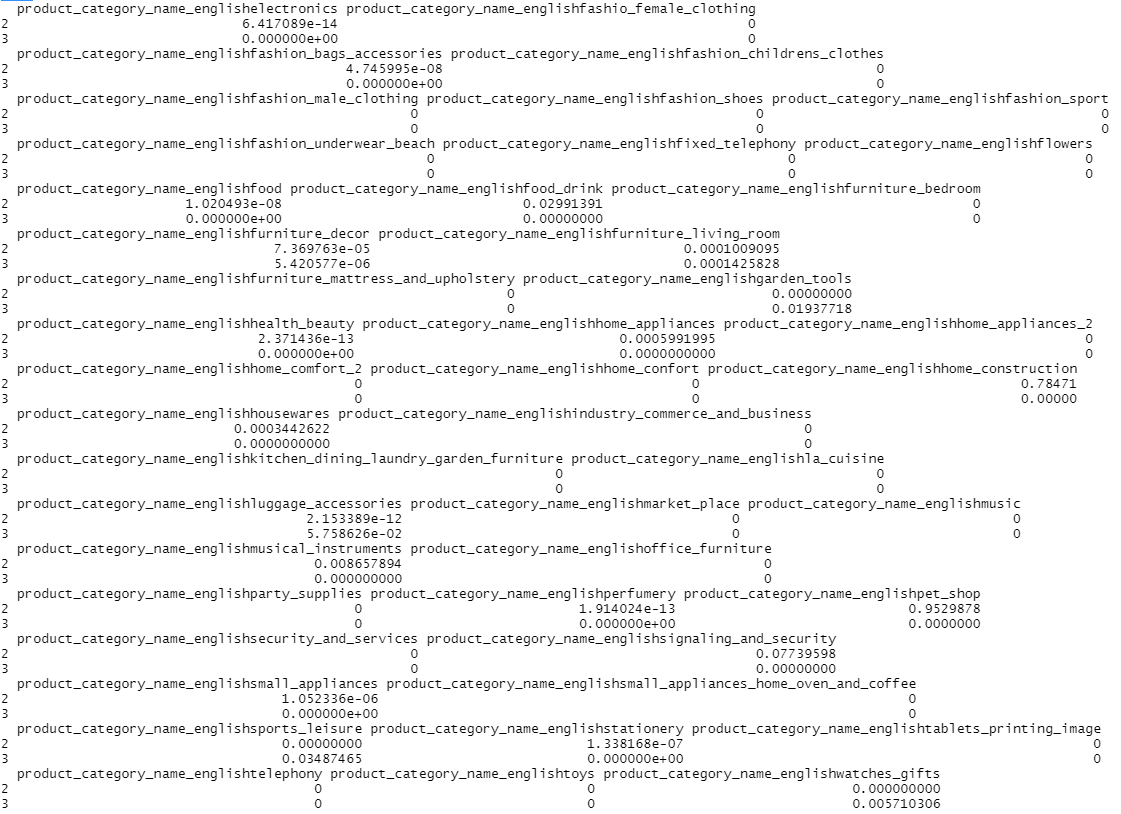


### G.3 P-Value of Variables for UCI Original Trainset



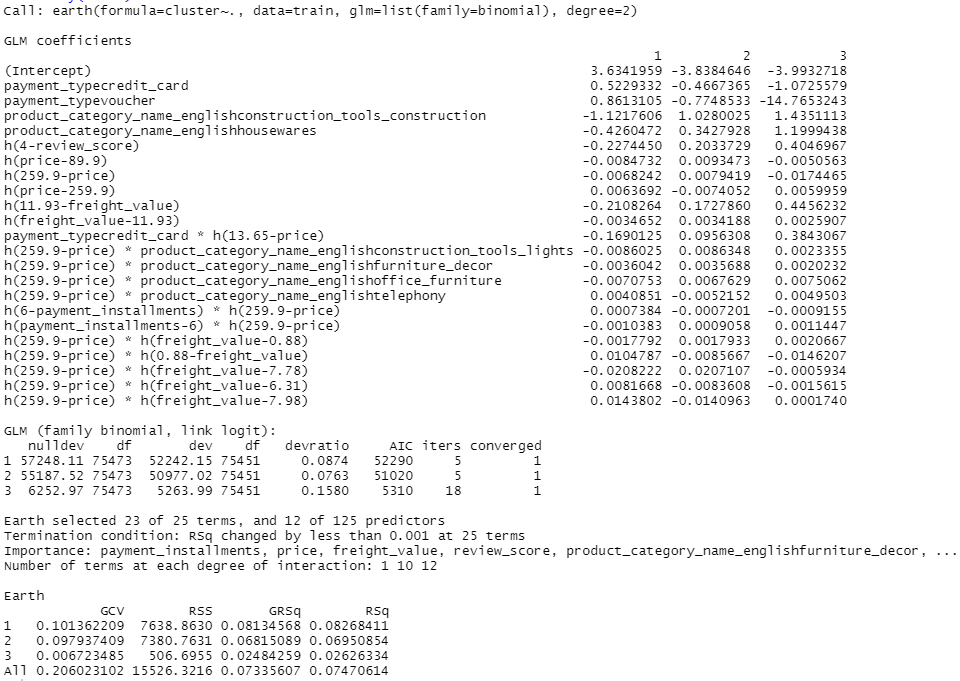
### G.4 P-Value of Variables for Olist Original Trainset



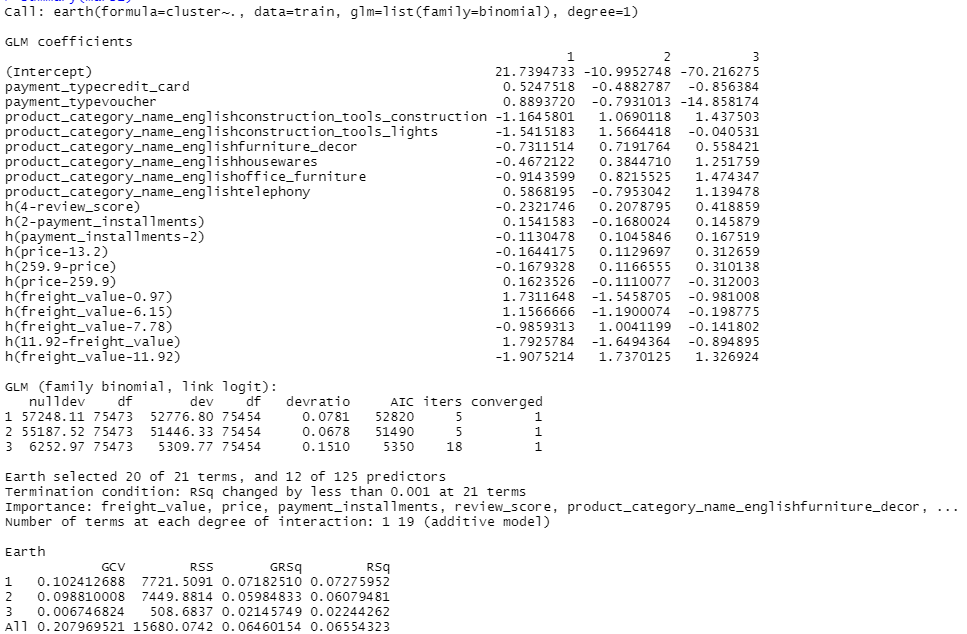


# Appendix H: MARS

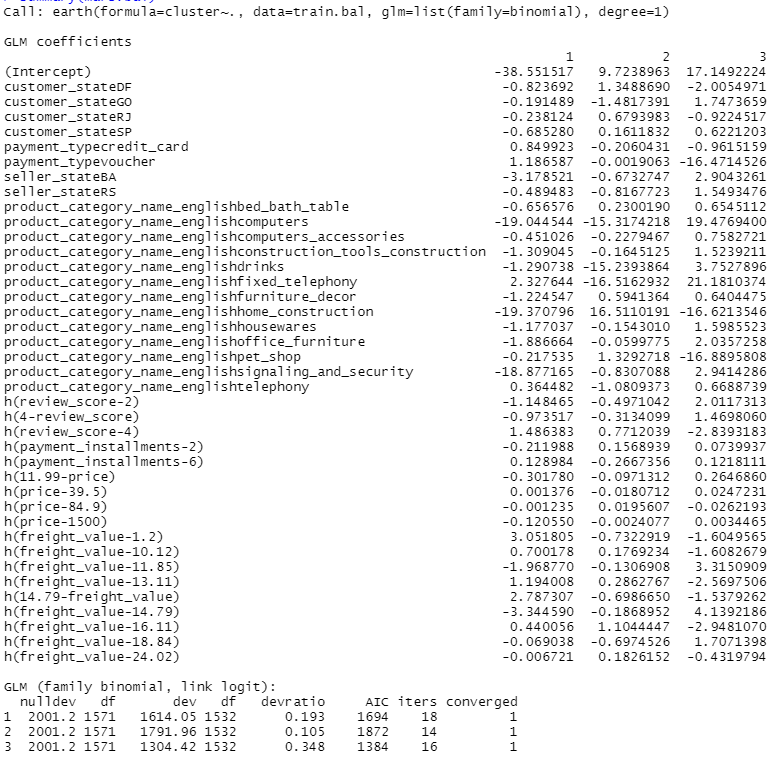
### H.1 Results of MARS Model with degree = 2 - Olist Original Trainset

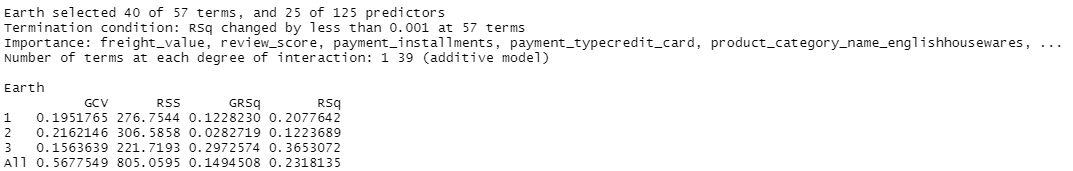


### H.2 Results of MARS Model with degree = 1 - Olist Original Trainset

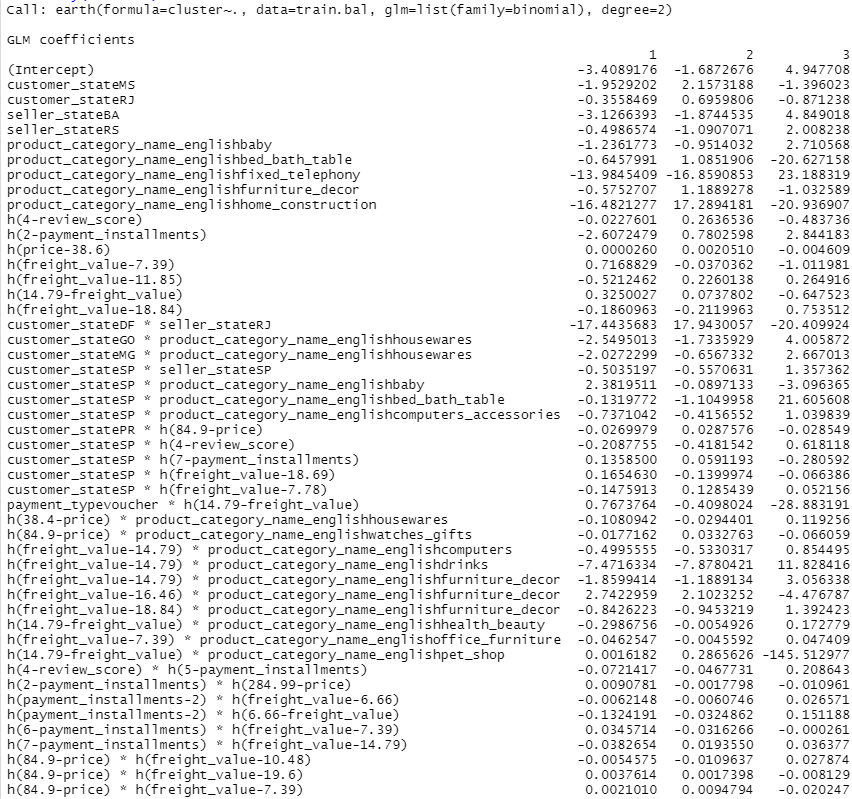


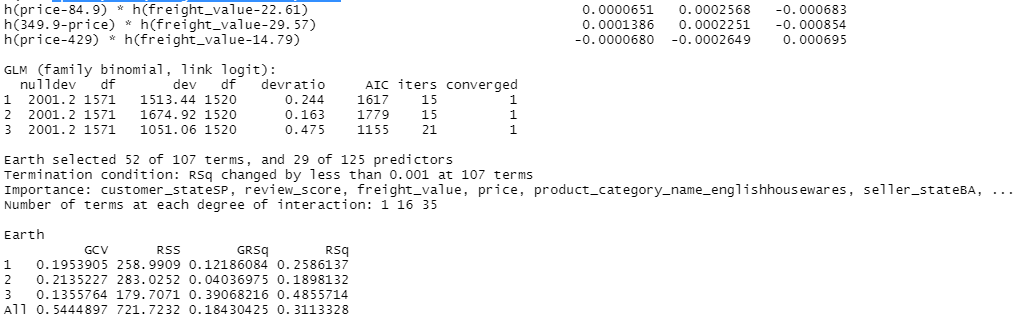
### H.3 Results of MARS Model with degree = 1 - Olist Balanced Trainset

****

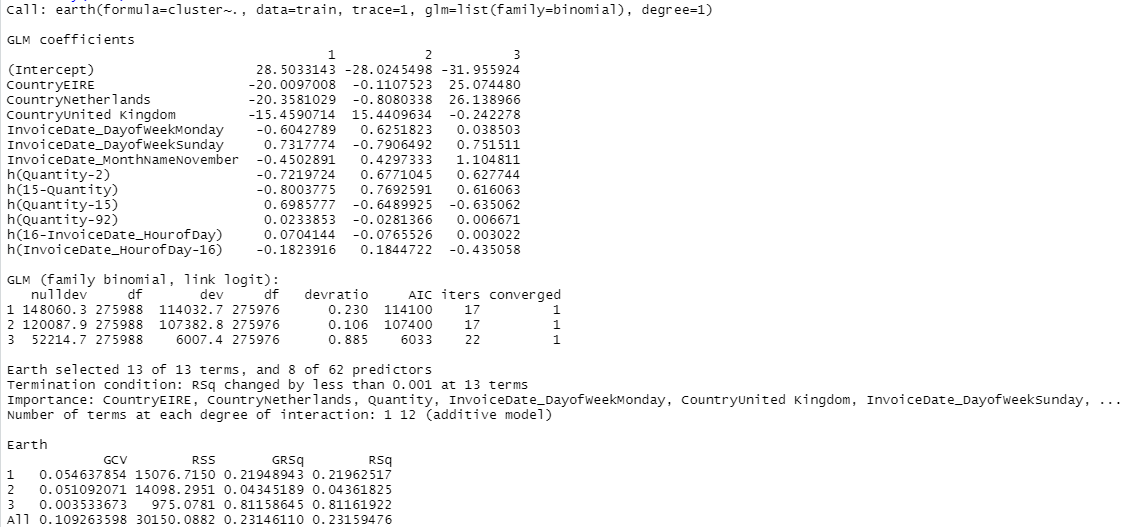
****

### H.4 Results of MARS Model with degree = 2 - Olist Balanced Trainset

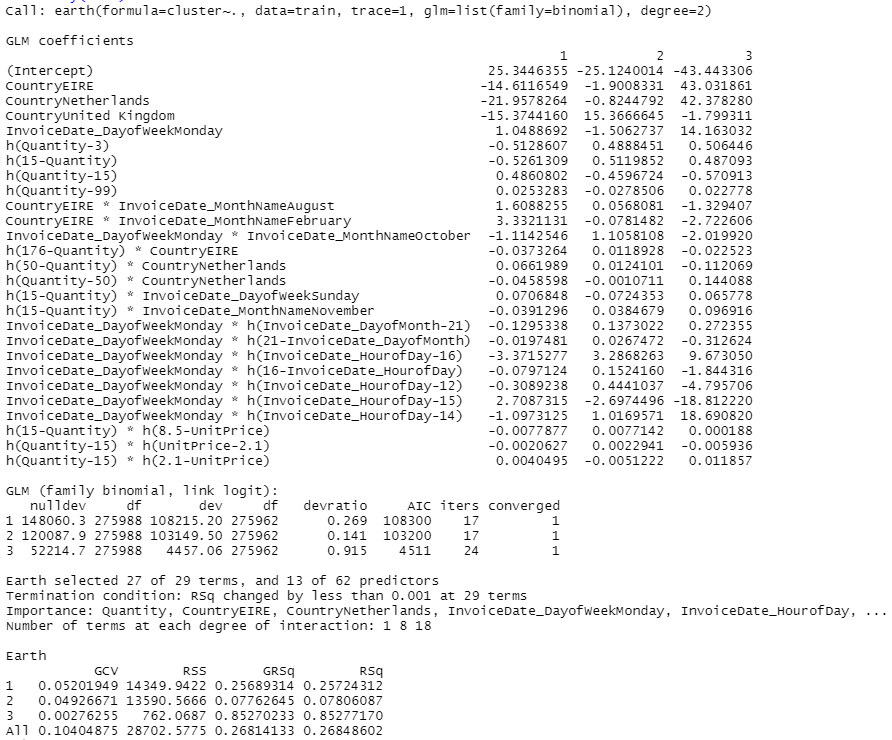




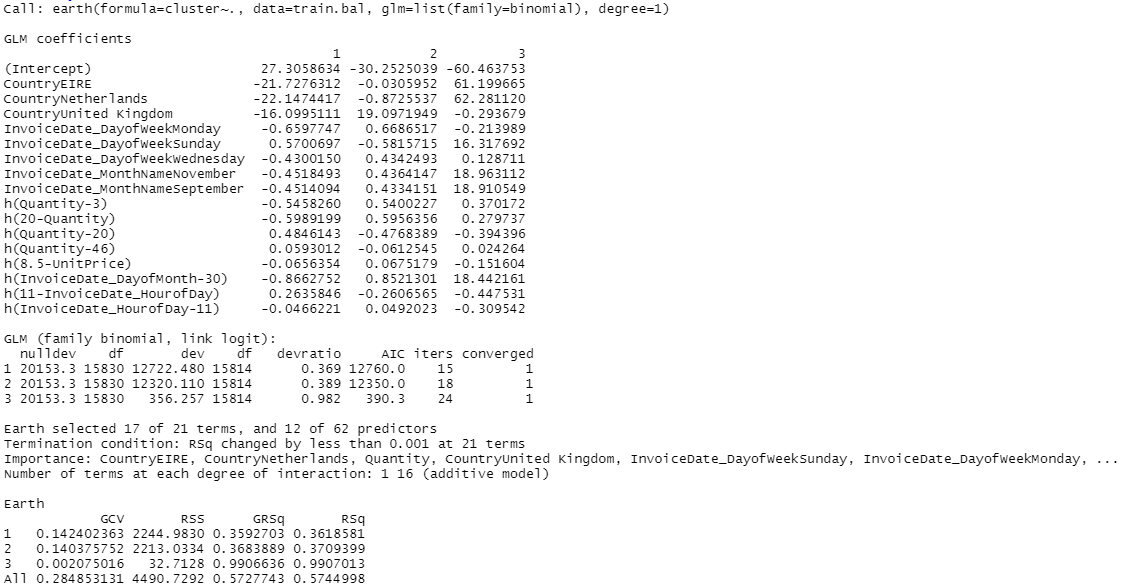
### H.5 Results of MARS Model with degree = 1 - UCI Original Trainset

****

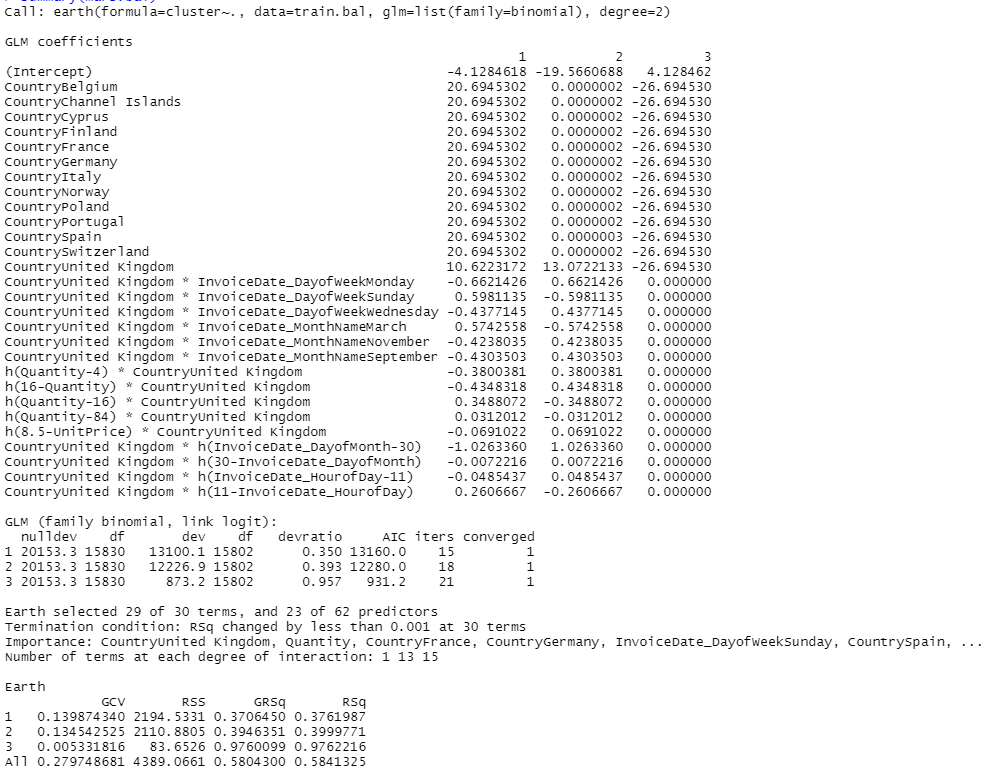
### H.6 Results of MARS Model with degree = 2 - UCI Original Trainset

****

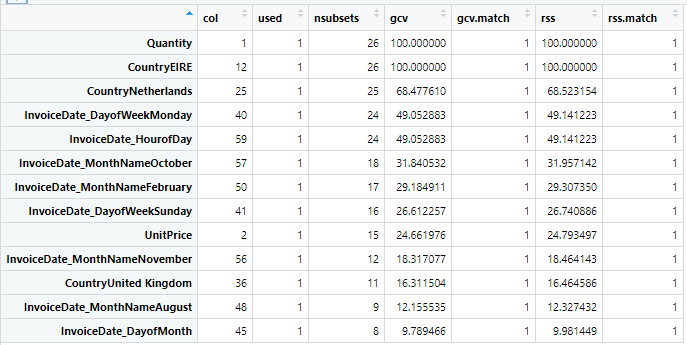
### H.7 Results of MARS Model with degree = 1 - UCI Balanced Trainset



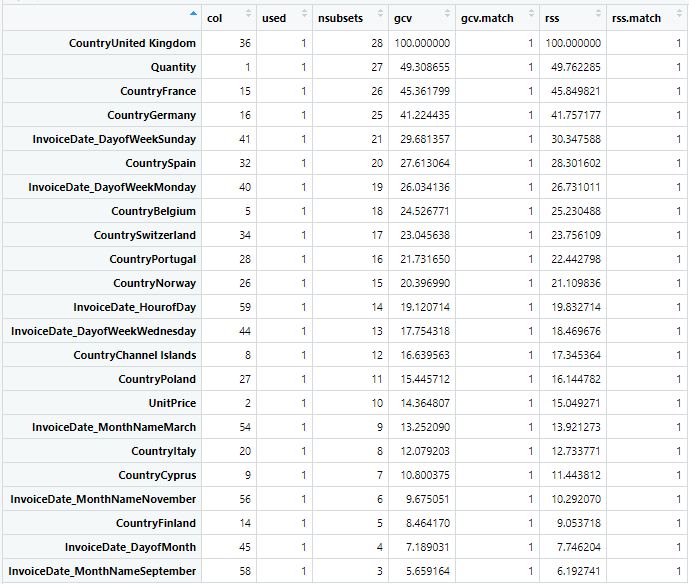
### H.8 Results of MARS Model with degree = 2 - UCI Balanced Trainset



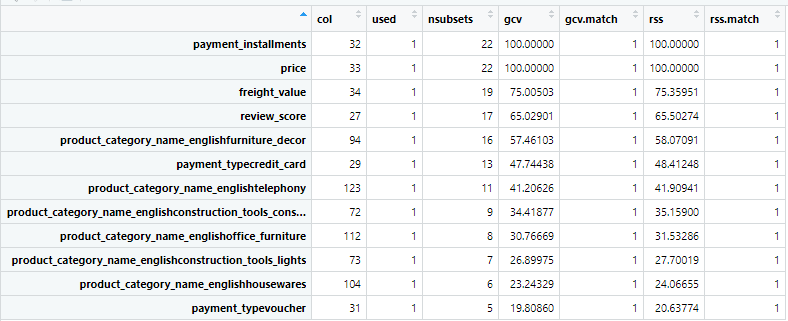
### H.9 Variable Importance generated by MARS Model - UCI Original Trainset



### H.10 Variable Importance generated by MARS Model - UCI Balanced Trainset



### H.11 Variable Importance generated by MARS Model - Olist Original Trainset



### H.12 Variable Importance generated by MARS Model - Olist Balanced Trainset

