Behavioral Modelling of Customer Marketing Patterns and Review Prediction Using Machine Learning Techniques

MSc Research Project
Data Analytics

Pratiksha Arvind Chate
Student ID: x20150377

School of Computing
National College of Ireland

Supervisor: Prof. Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Pratiksha Arvind Chate
Student ID:	x20150377
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Prof. Christian Horn
Submission Due Date:	31/01/2022
Project Title:	Behavioral Modelling of Customer Marketing Patterns and Review Prediction Using Machine Learning Techniques
Word Count:	6941
Page Count:	25

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

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

Signature:	Pratiksha Arvind Chate
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

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

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

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

Behavioral Modelling of Customer Marketing Patterns and Review Prediction Using Machine Learning Techniques

Pratiksha Arvind Chate
x20150377

Abstract

Technology has enabled businesses to produce a wide range of products that can dazzle clients, but customers are befuddled by the variety of choices available. While the retail industry faces several issues in terms of customer attention and loyalty, there is a need to improve marketing strategies. This study aims to segment the customers based on their purchasing patterns with RFM modelling and predict the review score for next order. The binary classification is performed to classify the reviews as positive (review score greater than 3) or negative (review score less than or equal to 3). Detailed analysis of the data have been performed on the Olist e-commerce dataset available publicly on Kaggle. New time-based and distance-based features were created from the existing attributes that were found to be useful for the prediction of review score. Machine learning classification models such as Random Forest, Light Gradient Boosting Model (LGBM) and AdaBoost model were implemented on the randomly oversampled data. The Random Forest Model outperformed other classification models with 95% accuracy and the F1-score of 0.95 for positive and negative class. This approach was found to be successful for detecting the review score and compared well with previous work in the field.

Keywords– AdaBoost, LGBM, Random Forest, RFM modelling, review prediction

1 Introduction

The advent of the internet caused a radical upheaval in business practises all over the globe, culminating in a drastic shift in marketing patterns that influenced both consumer and company behaviour. Furthermore, with the emergence of technological developments, a plethora of channels for communication have become available. The extensive use of digital services at all stages has created a slew of possibilities for both consumers and businesses. To stay competitive in today's fast-changing e-commerce market, a dynamic and customer-centric strategy is required. As a result, retail organisations spend extensively on adapting new flexible technologies in order to retain a competitive advantage in the market. Businesses must establish and implement clear customer-centric strategies for serving its customers as they are considered to be the foremost priority to retain its quality (Maryani et al.; 2018). The major emphasis of businesses is not on acquiring new prospective buyers, but on selling more items to current consumers, since the cost of acquiring new consumers is far higher than retaining existing ones (Tama; 2010).

As the business grows to become increasingly competitive, it becomes extremely essential for the firm to keep valuable and indispensable prospective clients (Chiliya et al.; 2009). It is imperative to have a more in-depth comprehension of each consumer, as well as their purchasing habits and preferences. Segregating clients into multiple groups based on their purchasing patterns and attributes is one of the most viable and optimal marketing strategies. This study makes an effort to segment customers using RFM modelling. RFM Modelling is a marketing paradigm that assists organisations to evaluate and examine behaviour of consumers based on 3 aspects, Recency, Frequency, and Monetary Value. This segmentation technique divides the customers into distinct homogeneous categories, allowing them to interact with different groups using distinct focused marketing approaches. In customer relationship management (CRM), the RFM model is an essential quantitative analytical model. The RFM model describes the significance and type of customer using three parameters: recency (R), frequency (F), and monetary value (M). Typically, enterprises utilise the RFM model and historical data to examine customer sales history and buying behaviour in order to discover future customers (Huang et al.; 2020). Additionally, customers assess the quality of product or service by the review score scaled from 1 to 5, 1 being the lowest and 5 being the highest. E-commerce sites grade products based on customer reviews and ratings, giving consumers valuable insight into how well a product performs. Nevertheless, from the seller's perspective, these evaluations will help improve the service offered. For the supplied historical data, this research seeks to forecast the review score as positive or negative for the customer's future purchase. This study attempts to explore and elucidate an answer the following research question using the pertinent approach indicated above:

How precisely does review prediction anticipate the customer feedback and standard of service for their future orders?

The objective of this research is to segment the customers based on RFM Analysis and further predict the review score for the historical data provided. The research uses the binary classification approach to predict the reviews using Machine Learning classification models such as Random Forest, LightGBM, and AdaBoost model. The implemented models are assessed based on the evaluation metrics such as accuracy and F1-score. The Figure 1 depicts the workflow followed for prediction of reviews.

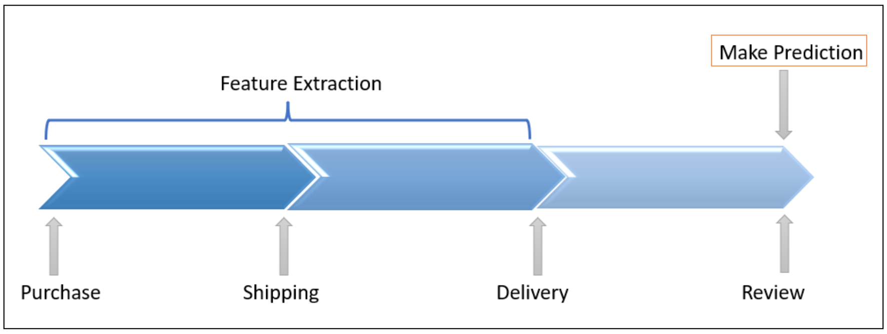


Figure 1: Review Prediction Workflow

The report is further divided into 6 different sections. Under Related Work in Section 2, a comprehensive critique of previous research is presented, highlighting the present stance on the selected area of interest. Methodology in Section 3 describes a succinct explanation of stages involved in implementing the methodology, which covers the data collection, data cleaning, exploratory data analysis, feature engineering, and rationale for picking the particular classification techniques. The underlying architecture is briefly described in the subsequent section under the Design Specification in Section 4. The next part discusses the Implementation and Model specifications in Section 5. The implementation's outputs are carefully evaluated for performed experiments/case studies, and the implications are briefly noted under Evaluation in Section 6. The study's ultimate judgement, as well as any potential future research that may broaden the scope of the research, is discussed in Section 7, Conclusion and Future Work.

2 Related Work

This area gives the summary of research papers related to enhancing client relationships, customer segmentation based on heterogeneity of features and purchasing behaviour, and the use of Data Mining for predicting the reviews rated by the customer. Additional information is provided in the subsequent sections.

2.1 Customer Segmentation

This section highlights the growing significance of understanding consumer behaviour and how it might help the organisation to maximise its profitability.

As business expands in size, it will be essential to split its consumers into an acceptable number of groups that are internally homogeneous and mutually heterogeneous based on certain commonalities in these consumers from the standpoint of marketing analysis (Hung and Tsai; 2008; Chang et al.; 2010). The characteristics of various segments may be compared, calculated and then analysed to offer important decisional information for administration. As a result, a variety of personalised marketing tactics may be employed to address a variety of customer requirements. Kotler and Keller (2006) determined that there are two categories of consumers factors such as consumer attributes and behavioral aspects. Customers are mainly characterised by geographical, demographical, and psychographical factors, whilst behavioural variables are comprised of attitudes about the products and the reactions consumers depict to the benefit, scenario, and brand (Wu and Pan; 2009). Using behavioural modelling, clients responsible for greater and lower earnings to a company may be targeted preferentially with relevant marketing tactics to maximise profit to stakeholders (Maryani et al.; 2018).

Mccarty and Hastak (2007) conducted an investigation using RFM, chi-square automatic interaction detection (CHAID), and logistic regression for direct marketing segmentation. In order to better grasp its potential as a database marketing analysis tool, a comparative study on these techniques was performed. RFM is regarded as a low-cost and typically dependable treatment with best results. The RFM (recency, frequency, and monetary) model is a behavior-based model used to assess a customer's activity and then create predictions based on the behaviour in the database (Yeh et al.; 2009). Furthermore, recency defines the length of time since the previous order, frequency represent the number of purchases inside a specific timeframe, and monetary value specifies the amount of money spent in this particular time period (Wang; 2010). Most segmentation methods utilise the RFM model. It consists of three measurements (recency, frequency, and monetary) that are integrated into a 3-digit RFM cell code that covers five equal quintiles (20 percent group). Recency is sometimes seen as the most essential of the three RFM metrics. Nevertheless, past research has shown that RFM levels are firm-specific and are reliant on the type of the goods (Lumsden et al.; 2008).

The RFM-based customer quantification procedure begins with sorting the information and separating consumers into 5 equal groups. Customers are categorised by purchase dates for recency. The more recent the date, the higher is its recency value. The highest 20% of the section is evaluated as 5, the next 20% as 4, and so on (Tsai and Chiu; 2004). Similarly, the highest quantile for frequency is awarded a value of 5 and the lowest performing is assigned a value of 1. A consumer with a higher frequency value indicates that customer has a strong desire for the product and is more likely to buy it again (Wei et al.; 2010). To minimise frequency and monetary co-linearity, Marcus (1998) proposed using the average buy amount rather than the total aggregate purchase amount. Ultimately, all consumers are shown by 555, 554,..., 111, resulting in 125 RFM cells, where segment 555 is considered to be the best performing customer segment and 111 as the least performing customer segment that needs attention (Wei et al.; 2010). When addressing the RFM model's scoring system, there are two approaches to generate an unique RFM value. One technique is to combine the recency, frequency, and monetary values collectively by combining the mean order and frequency each year, however the other way is more widely used in practise by summing the RFM values together (Miglautsch; 2000).

Decision makers can quickly grasp the implementation of the RFM model using RFM technique. (Mccarty and Hastak; 2007; Wang; 2010). RFM characteristics are retrieved using an internal data holding customer information about transaction history and are not obtained through aggregate level analysis in demographic data. As a result, RFM is more useful for targeting profitable business (Kaymak; 2001). Based on the literature reviewed, it can be discerned that RFM is a widely accepted data mining technique to segregate the customers. This research attempted to segment the consumers in 4 quantiles and combine the RFM values by summing up the three values.

2.2 Binary-Classification for Review Prediction

This section of the related work presents the approaches and techniques used in previous research carried out to predict the review score.

Reddy et al. (2017) proposed an approach to predict the star rating given by the customer based on the review comments. The classification models such as Naïve Bayes, Multinomial Naïve Bayes, Bigram Multinomial Naïve Bayes, Trigram Multinomial Bigram-Trigram Multinomial Naïve Bayes, and Random Forest were implemented and their performance was compared. It was found that the Random Forest classifier outperformed all the other models in terms of accuracy. Aravindan and Ekbal (2014) proposed a sentiment classification approach with feature extraction using Association Rule mining. WordNet, which offers details on the polarization of various words, was employed. Support Vector Machine was found to be the optimal and best performing model with 79.67% accuracy. Yu et al. (n.d.) implemented regression models such as Linear Regression and Random Forest regression model and Latent Factor model in order to predict the review score. On comparison of these models, Linear Regression and Latent Faction model were overfitting, however, random forest regression model achieved the best results without overfitting. Monett and Stolte (2016) conducted two sets of trials to estimate the star ratings of mobile applications based on the stated views from each review. One is concerned with determining the value of sentiment in reviews and filtering out reviews that lack sentiment. After examining the findings of the first set of trials, next round of studies employs different predictors such as text data. The highest results were obtained by filtering subjective words with neutral emotion and determining the total sentiment the review score. Pre-processing-porter Stemming, Part of speech tagging and stopping was implemented before feature extraction, impact analysis of features and feature reduction. Finally, categorization is accomplished via the use of Bagging, Random Forest, Decision Tree, Naive Bayes, and K-Nearest Neighbor. Random Forest performed the best in terms of accuracy, while Naive Bayes had the lowest accuracy.

Leveraging non-respondent data in customer satisfaction modeling (2021) implemented unique method for identifying data-driven qualities that accurately describe consumer interactions to forecast consumer satisfaction. The application of Artificial Neural Networks is utilized to predict customer satisfaction in banks (Zeinalizadeh et al.; 2015). Supervised machine learning classification algorithms were employed to experimentally identify key variables and subsequently categorise satisfaction of customers for Las Vegas hotels based on Yelp reviews, the results of which demonstrated that supervised learning classifiers provide more trustworthy results than previous statistical techniques (Sánchez-Franco et al.; 2019).

The use of a Random Forest classifier has been shown to yield improved accuracy in comparable classification challenges. It was also discovered that boosting methods increased model performance when compared to standard supervised classification models (Dietterich; 2004). This research is based on the implementation of binary classification problem for predicting the review score as ”positive” or ”negative” using classification models such as Random Forest, Light Gradient Boosting Model (LGBM) and Adaptive Boosting (AdaBoost) model and compare the performance of these models to find the best one. These implemented models are assessed based on the evaluation metrics such as Accuracy and F1-score as measures like accuracy and the F1 score are often employed to assess classifier performance in machine learning (Wardhani et al.; 2019).

3 Methodology

Methodology section incorporates the detailed overview of intended methodological approaches. This research aims to follow the Knowledge Discovery Database (KDD) approach as depicted in the Figure 2. The desired strategy entails following tasks:

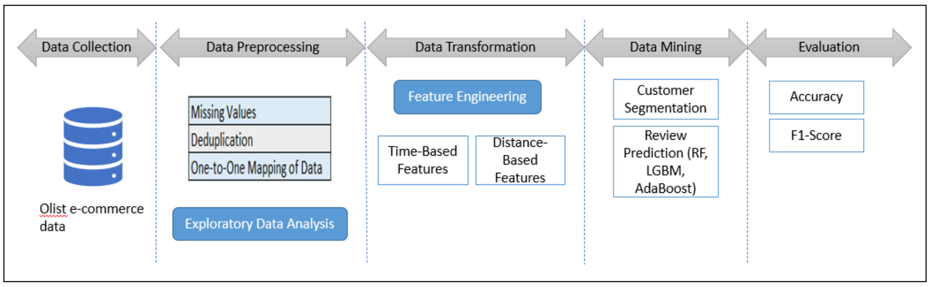


Figure 2: KDD approach for review classification

3.1 Data Collection

The dataset which the research is based on is a fairly descriptive sales and order data of an e-commerce marketplace integrator, Olist, founded in 2015 in Brazil. The dataset incorporates particulars of 100k customer orders placed in Brazil between the year 2016 and the year 2018. The attributes of this data facilitate observation of details from several viewpoints, such as product parameters, online purchases, sellers, and customers. A location specific data relating Brazilian zip codes to latitude/longitude coordinates are made publicly available. This data in the form of CSV is further subdivided into numerous distinct datasets for easier interpretation and organization as depicted in the Figure 3. The authenticity of the data will help in producing and analyzing factors accurately that affect the customer behavior and their corresponding purchase patterns. Moreover, the data is structured and normalized, which will result in optimized computational time for processing. The details encapsulated within different CSV files within this dataset hold significance with respect to each order such as customers, payments, items, reviews, orders, products, sellers, and so on.

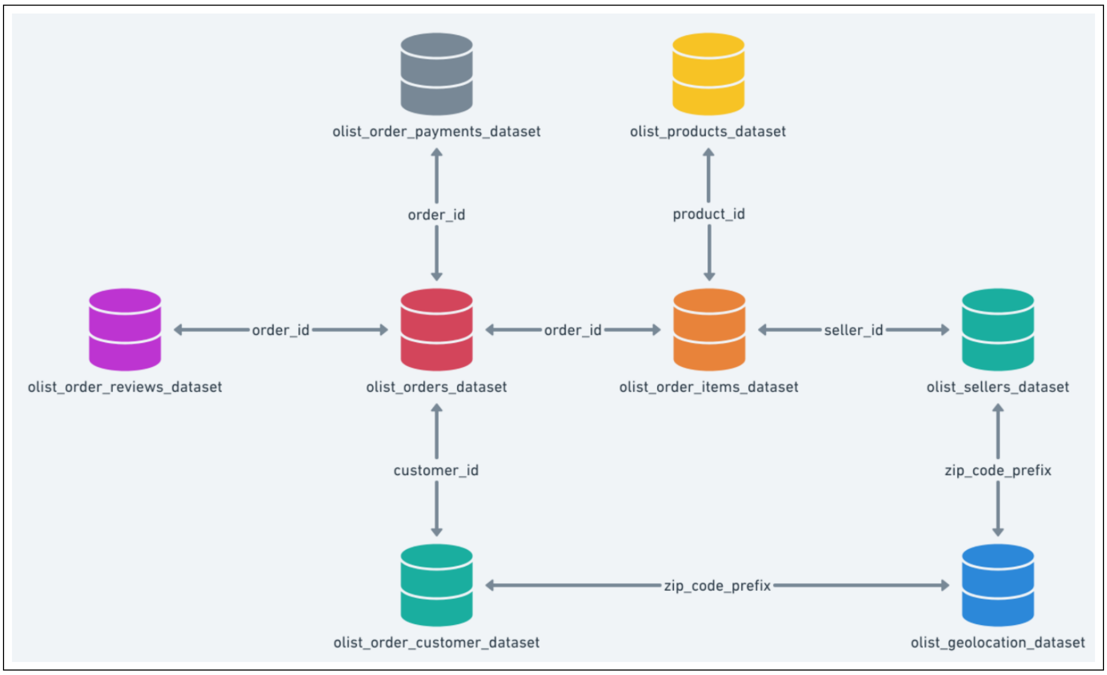


Figure 3: Structure of the Dataset
Source : https://www.kaggle.com/olistbr/brazilian-ecommerce

The orders data table contains specifics about order for each online purchase that are linked to other tables with customerID and orderID. The reviews table contains the categorized review of products for unique orders in the order table. The categories of reviews are scaled from 1 to 5 where 1 is considered to be the least rating and 5 as the highest rating. As per the analysis and interpretation of this study, it perceives rating above 3 as a positive review i.e., the product is well received by the customers whereas a rating equal to or below 3 is considered as a negative review i.e., the quality of the product or service does not meet the expectations of the receiver. Also, this dataset is publicly available for research purpose and will not hold any ethical implications.

3.2 Data Preprocessing

3.2.1 Data Cleaning

Data cleaning attempts to address the issues related to data discovered during data gathering process. The extracted datasets were merged to generate a single data frame for detailed data analysis. However, following discrepancies were observed:

Discrepancies:

- While integrating these datasets, there was a discrepancy observed in the Reviews data table where, one order ID was tagged to multiple reviews eventhough there was only one product or item purchased.
- A common review ID was tagged to multiple order IDs.
- Another similar scenario was observed in an event where, one order ID has multiple order item IDs and payment IDs.

The above discrepancies result in extrapolation and cartesian values which might result in incorrect computations. In order to overcome these discrepancies, order items, reviews and payment tables were analyzed to check for the uniqueness of Order ID variables across these tables. The percentage of redundant Order IDs across Order Items, Reviews and Payments was found to 9.97%, 0.55% and 2.97% respectively. Since, the proportion of redundant data was miniscule, such Order IDs were dropped from the dataset to adhere to the assumptions formulated below.

Assumptions:

- Order ID will be unique across all the transaction tables so that the resultant data is based on one-to-one mapping.
- One review ID can be tagged to just one order ID.
- For analysis, the data will be constrained to contain unique order IDs with one order item and one payment record.

The final dataset combined for further data analysis consists of null values and duplicate values in some columns. The entries review title and review message in the Reviews data table have the highest number of missing values whereas most of the columns in the Products table has few missing values.

The missing values are treated by replacing them with similar values or simply by dropping the less significant attributes. Some of the timestamps such as order approved date are replaced by order purchase date and similarly null values in the order delivery date are replaced by estimated delivery date. The column order delivered carrier date have been dropped. Other less significant columns such as review ID, review comment title, review comment message, review creation date and answer timestamp have also been dropped as these columns will not be a part of analysis. The missing values in the numerical features such as product weight in grams, product length, product width, and product height in centimeters have been replaced with the median of the respective columns and the null values in the product photos quantity have been dropped.

3.2.2 Exploratory Data Analysis

The large number of available attributes post data cleaning encourages the scope for a detailed analysis about the marketing patterns of customer as well as sales performance of the organization. The research performs an extricate study of review analysis in conjugation with relative factors such as Product Categories, Payment Channels, Delivery time, Delivered products and more. In order to improve the service and quality of products, the positive and negative review comparison is carried out with respect to different product categories, states, Month of the Year, day of the week and time of the day, products (delivered, shipped, etc.), delivery time etc. In order to check statistical significance a set of chi-square tests were performed between variables such as payment type. The data constitutes for the time frame between 2016-2018 which urges to analyze the trend of some temporal features such as Monthly Sales, Order Volume, Customer Distribution across all 3 years.

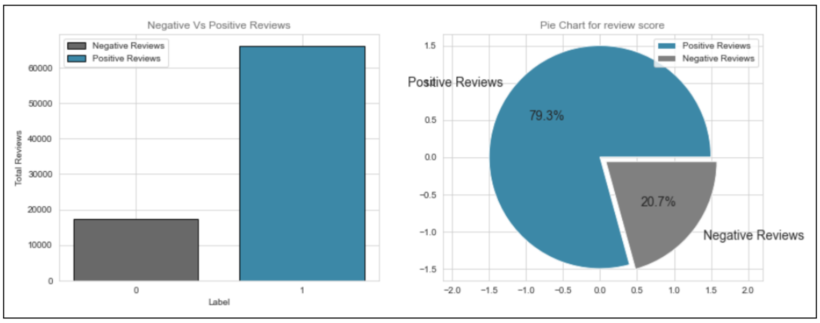
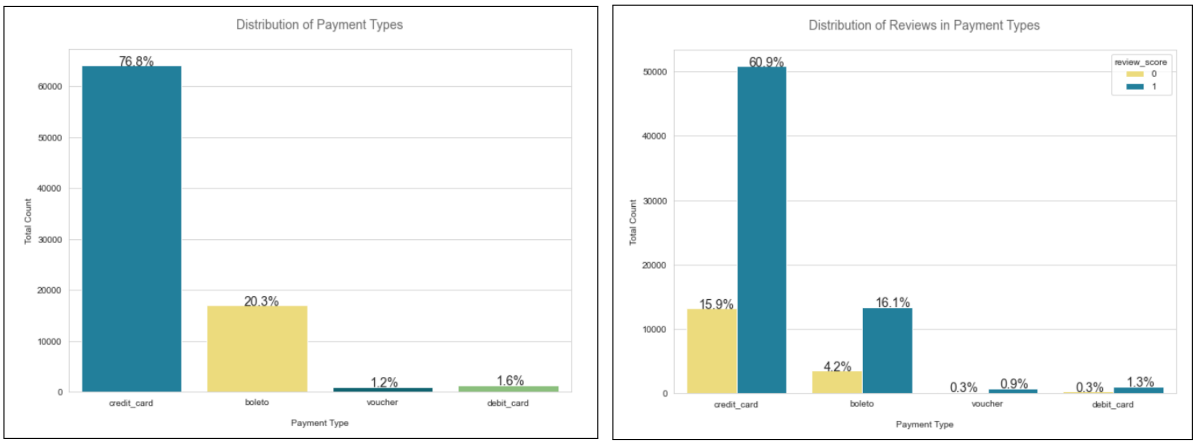


Figure 4: Distribution of Positive and Negative Reviews

It is observed that the dataset is highly imbalanced (refer Figure 4). Although, this data is organically authentic and non-fictional. In Figure 4, a bar chart and a pie chart illustrates the distribution of positive and negative reviews. The dataset comprises of nearly 79% positive and around 21% negative reviews. The same data is also conveyed through the pie chart.



(a) Distribution of Payment Types

(b) Reviews Distribution in Payment Types

Figure 5: Payment Type

From the above plots in Figure 5a, it can be inferred that the most used payment method is credit card with 76.8% followed by boleto. Boleto is a nationally recognized payment method in Brazil accredited by Brazilian Federation of Banks (20.3%). Credit card and Boleto mode of payment is responsible for the majority of the transactions (approx 97%). From Figure 5b, it can be clearly observed that most of the customers have given positive reviews irrespective of the mode of payment. A chi-square test with significance level, alpha=0.05 is performed to check if payment type and review score are dependent with null and alternative hypothesis as stated below..

- **Null hypothesis:** Payment type and review score are independent.
- **Alternative hypothesis:** Payment type and review score are dependent.

As a result, the null hypothesis was rejected with the p-value of 0.00018. This test concluded that review score is dependent on payment type and this dependence is statistically significant.

From the detailed analysis of data implemented, it is observed that maximum customers are from São Paulo constituting 42% of the total customers followed by Rio de Janeiro (approx 13%) and Minas Gerais (around 12%).

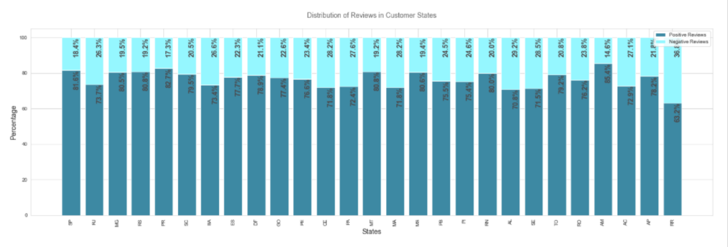


Figure 6: Statewise Distribution of Reviews

From stack plot of reviews by state in Figure 6, it may deduced that the majority of consumers in each state have submitted positive ratings. It is found that the maximum positive reviews i.e., 85.4% were from the state Amazonas(AM) followed by São Paulo(SP).

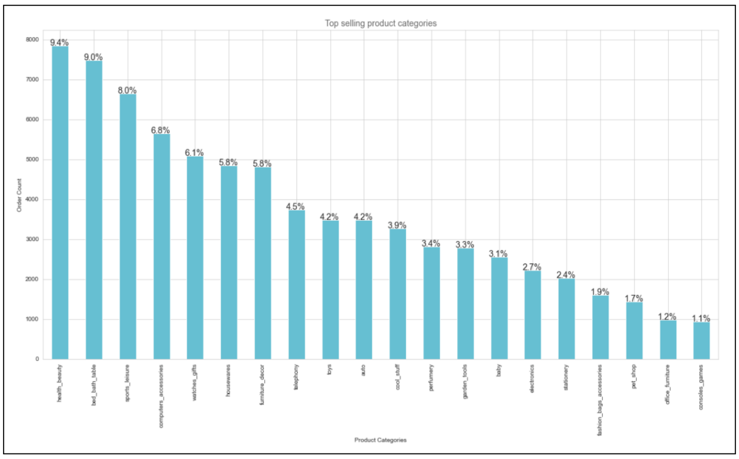


Figure 7: Top selling product categories

Figure 7 reveals the top 20 selling product categories of Olist store. During 2016 and 2018, the far more purchased items were in the health beauty(9.4%), bed bath table(9%), and sports leisure categories(8%), however, fashion children clothing and Security and services goods received the fewest orders.

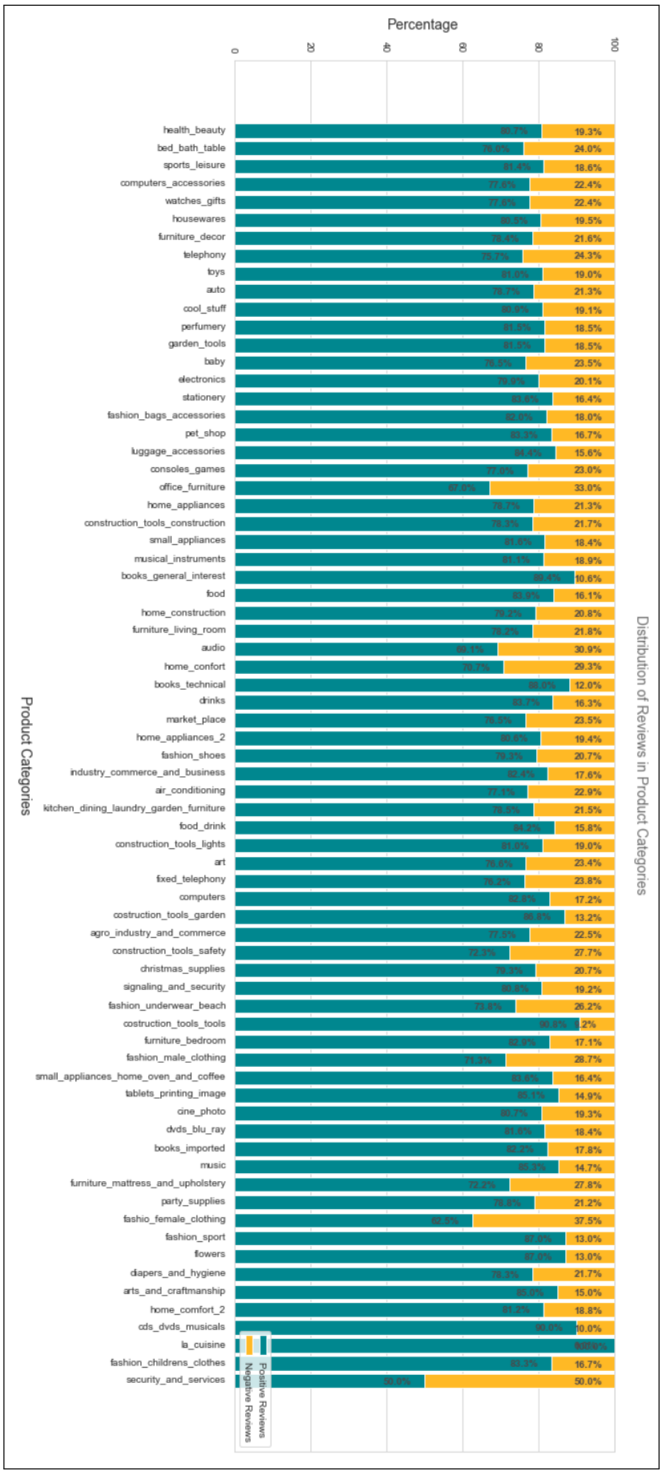


Figure 8: Distribution of Reviews in Product Categories

The stacked plot in Figure 8 displays the percentage of positive and negative reviews in all the product categories. Based upon the graph, it can be deduced that the majority of responses for the top selling product categories are favorable, and the same holds true for the other product categories. ”La Cuisine” is one particular product category where all the ratings are positive. This could be considered as the qualitative product category.

According to the analysis performed on order status, almost 98% of the orders have been delivered successfully, and 1% of the orders have been shipped for delivery. However, nominal orders (around 0.3%) have either been invoiced, cancelled, or are in processing. Also, 79% of the delivered orders have received positive ratings. However, 19% of the reviews for delivered orders are negative. It can be observed that the majority of the shipped orders that were not delivered earned negative ratings. From this analysis, it can be inferred that the delivery time of an order might be a critical aspect in solving this problem.

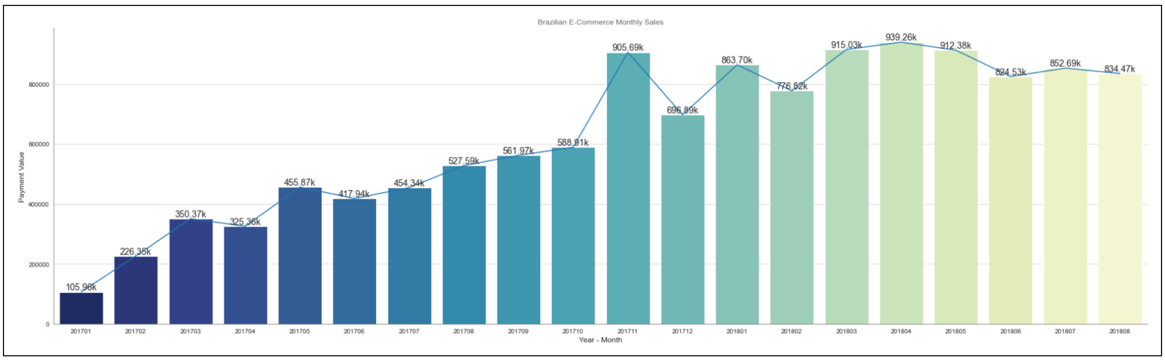
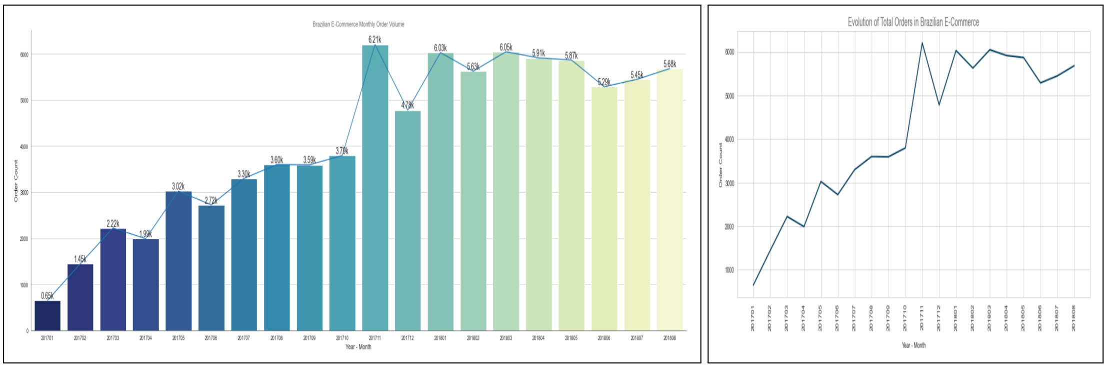


Figure 9: Monthly Payment Value of Olist

The dataset consisted of order data from September 2016 to September 2018. However, there was no data present for the month of November 2016 and the overall data for 2016 constituted barely 0.4%. In addition, there was just one record for September 2018. As a result, this data has been discarded. The Figure 9 shows the monthly payment value of the Olist store from January 2017 to August 2018. The revenue generated by Olist increased considerably till October 2017, with slight fluctuations. In November 2017, the highest payment value (905.69k) was reported, with a substantial increase. However, a downturn with mercurial variations was noticed following the month of November.



(a) Monthly Sales Order Volume

(b) Evolution of Total Monthly Orders

Figure 10: Evolution of Order Volume

The Figure 10 shows the evolution of monthly order volume in Brazil from January 2017 to August 2018. It can be observed that order volume grew significantly in the given timeframe. In the month of November 2017, there was a considerable increase in the monthly sales where it can be speculated that costs were perhaps lowered and discounts and consessions were offered owing to Black Friday. However, a decline was observed in the number of orders in the year 2018 and has since been fluctuating. There might only be two plausible causes for the sales decline that happened across most areas in 2018.

1. Competitiors in Brazil execute an equivalent campaign to entice users and purchases to their websites.

2. There might possibly be several severe flaws and adjustments to the business approach.

The very first assumption was considered as the most credible because the scenario anticipated no serious defects and adjustments in the business strategies. The hypothesis is further supported by the evidence[1]. It confirms that by October 2017, Amazon announced the very first strategic foray into commercial activity when it allowed third-party vendors to use its Brazilian online platform for their business.

From the analysis performed, it was also inferred that the month of August has the greatest percentage of favorable ratings across all reviews from January to December, with approximately 9.3% of the positive ratings. Positive evaluations are also more prevalent in the months of April and May with nearly 8.6%. The least number of positive ratings was observed in the month of October with only 3.5% of the positive reviews. However, July recorded the maximum number of negative reviews (2.8%). Also, it was discovered that most of the purchases are processed in midday, and perhaps the maximum number of positive feedback i.e., approximately 30% are submitted at that time. Also, order volume is large on Mondays with highest number of good ratings i.e., 12.9% followed by Tuesdays (12.7%).

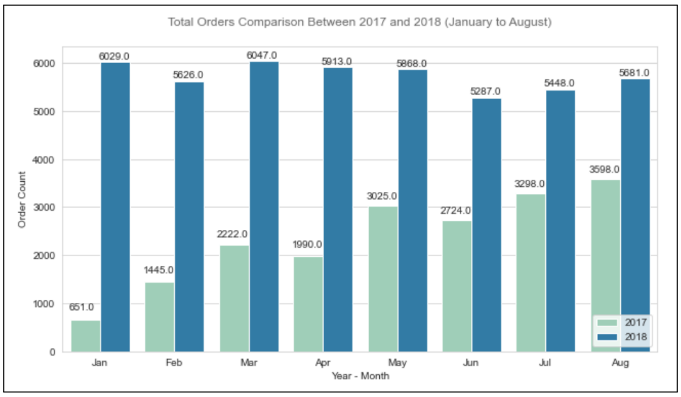


Figure 11: Comparison of Sales in 2017 and 2018

The Figure 11 illustrates the comparison of sales in 2017 and 2018. However, the two-year comparative analysis of sales is performed for the period of January to August due to unavailability of data for the remainder of the year, i.e., September, October, November and December for 2018. It is observed that the sales of the year 2018 were significantly higher for every month than the sales of 2017.

[1] Amazon's move in Brazil: `https://www.reuters.com/article/us-amazon-com-brazil/amazon-com-starts-direct-sales-of-merchandise-in-brazil-after-delays-idUSKCN1PG0AG`

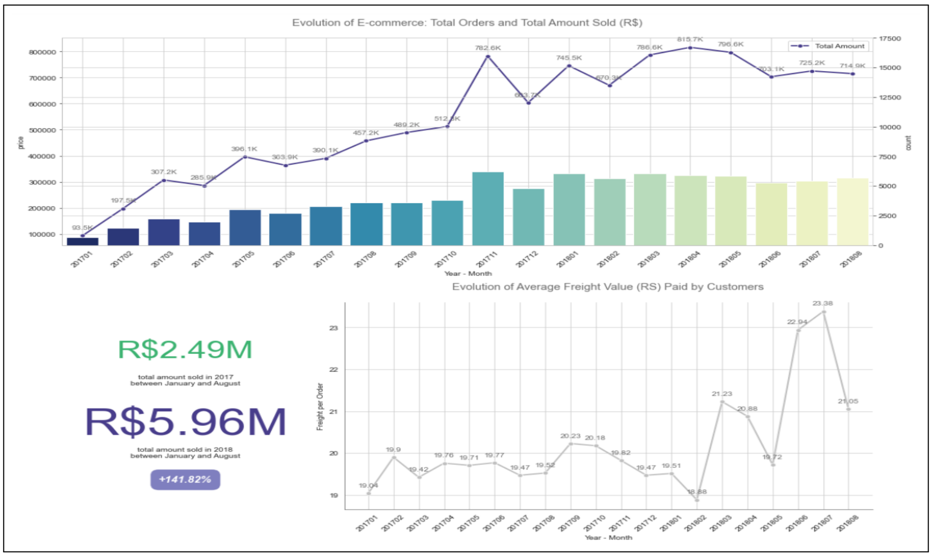


Figure 12: Evolution of Total Orders, Total Amount, and Average Freight Value

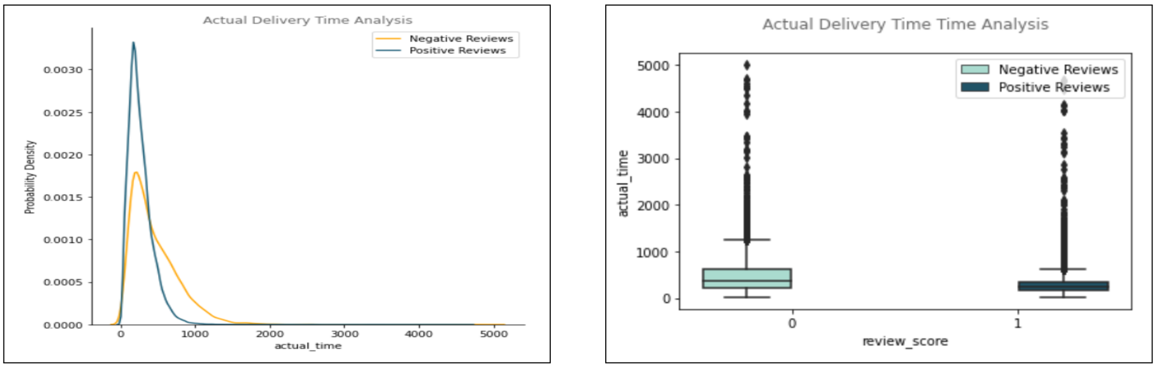
The above Figure 12 is a snapshot summary of total number of orders and total sales amount distribution across the period of January 2017-August 2018. The line plot on the bottom right depicts the average freight value paid by customers for the same period. The number of orders and sales since the start of 2017 witnesses an upward trend till November 2017 implying the exponential growth. However, these metrics fluctuate for the rest the timeframe till August 2018. The evolution of average freight value can be observed in the year 2018 with major shifts in the month of March, June, and July. A numerical metric on the bottom left delineates the upsurge in the sales amount sold in 2017 and 2018.

3.3 Data Transformation

Data Transformation or feature engineering may enhance the performance of the machine learning model significantly. Especially in this study, the existing set of attributes are not correlated with the target variable (review score). For effective data modelling, the data is further explored to extract some insightful information and new features are introduced from the existing features.

3.3.1 Time-Based Features

Delivery time is regarded as the most crucial aspect in e-commerce businesses. If the order is not delivered in the specified timeframe or if the expected delivery time is too lengthy, there is a good risk of customer dissatisfaction. However, if the estimated delivery time is shorter or the product is delivered at the promised time, the consumer is more likely to have a pleasant experience. Time-based features such as ”estimated time”, ”actual delivery time”, ”difference between actual and estimated delivery time”, ”difference between purchased and order approved time”, and ”difference between purchased and shipped time” have been created from the existing features to check if these are correlated with the target variable.



(a) Review Analysis for Actual Delivery Time

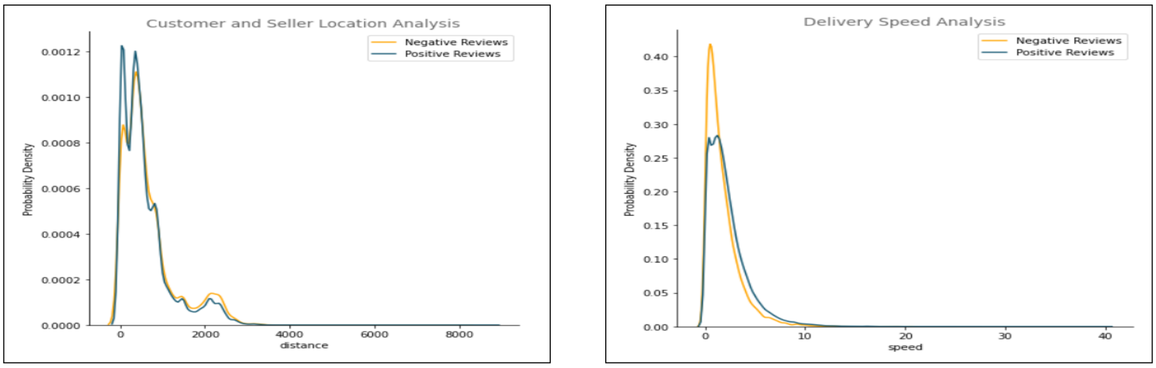
(b) Box Plot representation

Figure 13: Actual Delivery Time Analysis

From Figure 13 illustrates that there is a high risk of getting negative rating if the delivery time is long. From Figure 13a, the probability density function of the positive ratings is peaked for lower values of actual delivery time taken. When compared to the probablity density of negative review ratings, the probability density of positive review scores falls extremely sharply to the right. From the boxplot (refer Figure 13b), it is observed that the position of the boxplot is shifting upwards from a positive to a negative review score. Despite the lack of complete separation, this aspect is crucial for categorising review ratings. Though there are outliers present, this observation holds true for percentiles (50th, 75th, 90th) that were analyzed. This indicates that if the delivery time is longer, the product is more likely to get negative review score. However, if the delivery time is short, the odds of receiving a good review are higher. The similar behaviour is seen for ”difference between actual and estimated time”, where, the probability density function for positive reviews is highly peaked and this could be considered as an important feature to categorise the review score.

3.3.2 Distance-Based Features

It was observed from the EDA that most of the customer and seller base of Olist is from Sau Paulo state with majority of the favourable ratings. The distance between consumers and sellers could be considered as one factor that may contribute to quicker delivery times and, as a consequence, higher customer satisfaction. The new feature distance between the customer location and seller location (in km) has been created followed by the new feature speed of delivery.



(a) Review Analysis for Distance

(b) Review Analysis for Speed

Figure 14: Review Analysis for distance between customers and sellers and delivery speed

The Figure 14 demonstrates the categorization of reviews in the data for distance between customers and sellers and delivery speed. From Figure 14a, it can be clearly inferred that for shorter distance between customers and sellers, the positive review score is highly peaked. Negative review ratings have a larger density than other scores at increasing distances. From Figure 14b there is a high chance of getting a negative review score if delivery speed is low.

3.3.3 Binary Features

Binary features such as same city is created to check if the customer and seller are from the same state. Moreover, other binary features such as late shipping and high freight have been created.

3.4 Data Mining

3.4.1 Customer Segmentation

Customer acquisition is significant for any business but customer retention is even more critical, the reason primarily being the loss of entire series of purchases the customer would make over the course of its lifetime as a consequence (Wei et al.; 2010). As derived from the literature review, RFM analysis has been proven to yield better results as compared to other data mining techniques for segmentation of customers. Applying RFM analysis, the recency, frequency, and monetary value for each customer was evaluated. Post feature extraction, a quintile approach was implemented to divide customers into 4 quartiles resulting in 64 (4x4x4) RFM cells, where 444 is considered as the best performing segment and 111 as the least performing segment. RFM score was then derived by adding the recency, frequency and monetary quintile values together. The consumers were then segmented into different categories based on the RFM score as illustrated in Table 1

RFM Score	**Characteristics**
RFM Score ≥ 9	Can't Loose Them
$8 \leq$ RFM Score < 9	Champions
$7 \leq$ RFM Score < 8	Loyal
$6 \leq$ RFM Score < 7	Needs Attention
$5 \leq$ RFM Score < 6	Potential
$4 \leq$ RFM Score < 5	Promising
RFM Score < 4	Require Activation (i.e. RFM score 111)

Table 1: Segmentation based on RFM Score

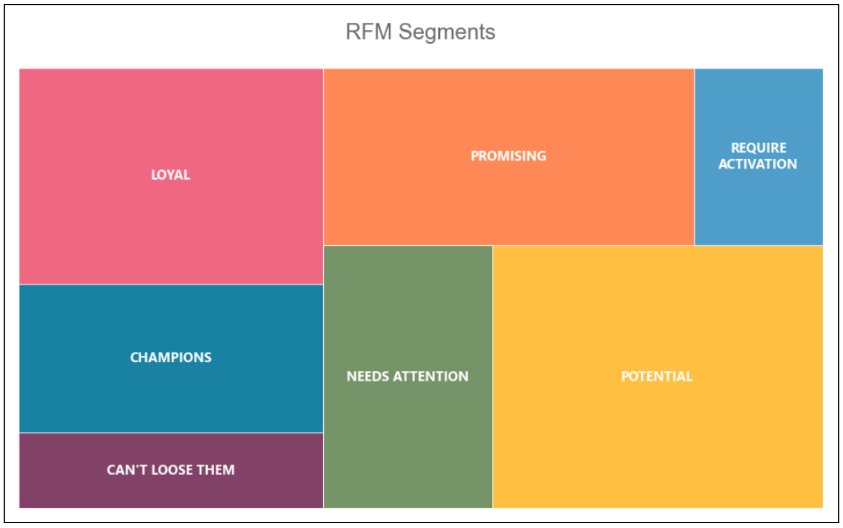


Figure 15: Customer Segmentation based on RFM score

From the Figure 15, it is clearly evident that most of the customers are in the "Potential" segment followed by "Promising" and "Loyal" segments.

3.4.2 Binary Classification for Review Prediction

This study is based on the binary classification in order to predict the review score as positive (review score > 3) or negative (review score ≤ 3). In order to achieve this objective, the machine learning classification models such as Random Forest (RF), Light Gradient Boosting Model (LGBM), and Adaptive Boosting (AdaBoost) are used based on the literature review.

1. **Random Forest (RF):**
   The random forest learning approach is made up of n sets of de-correlated decision trees. It is based on the notion of bootstrap aggregation, which is a strategy for rescaling with replacement to reduce variance (Kirasich; 2018). In order to achieve the best results, multiple experiments with the Random Forest classifier are carried out, each with a different approach.

2. **Light Gradient Boosting Model (LGBM):**
   The another classification model used to predict the review score is LightGBM, also

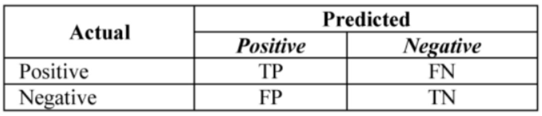
known as the parallel voting Decision Tree (DT) technique, uses a histogram-based strategy to expedite training, reduce memory consumption, and combine advanced network connections to maximise parallel learning. Furthermore, LightGBM develops trees leaf by leaf, aiming for the leaf with the largest rise in variance to split (Machado et al.; 2019).

3. **Adaptive Boosting (AdaBoost):**
   An AdaBoost classifier is a meta-estimator that begins by training a classifier on the original data and then fits further copies of the classifier on the same dataset with the weights of incorrectly identified cases altered such that subsequent classifiers focus more on challenging scenarios.[2]

3.5 Evaluation

Performance metrics such as Confusion Matrix, Accuracy, and F1-score were used to assess the machine learning classification models implemented.

1. **Confusion Matrix:** The confusion matrix is used to compute the metrics. Figure 16 depicts a confusion matrix for binary classification. The terms TP (true positive) and TN (true negative) indicated the number of samples properly categorised in the positive and negative classes, respectively. The numbers FN (false negative) and FP (false positive) indicated the number of samples in the positive and negative classes that were erroneously categorised (Wardhani et al.; 2019)



Actual	**Predicted**	
	Positive	***Negative***
Positive	TP	FN
Negative	FP	TN

Figure 16: Confusion Matrix
Source : https://ieeexplore.ieee.org/document/8949568

2. **Accuracy:** The classification accuracy is defined as the ratio of accurate predictions to total input samples.[3] In this case, the accuracy might be measured by taking the ratio of correct predictions and dividing it by the total input samples. The accuracy is formulated as mentioned in eq(1) below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

3. **F1-score:** The F1 score was calculated as the harmonic mean of precision and recall and is formulated as given in eq(2) (Wardhani et al.; 2019).

$$\text{F1} = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{2}$$

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html

[3] https://www.sciencedirect.com/topics/engineering/classification-accuracy

4 Design Specification

The implemented research follows a two-tier approach with data persistence tier and business logic tier. The Figure 17 below displays the two-tier architecture used to implement this research briefly outlining the methods followed as well as the tools and technologies used.

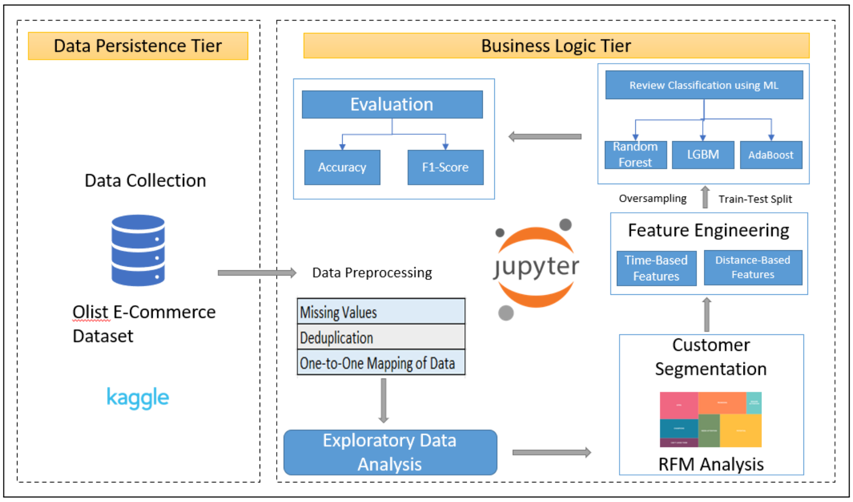


Figure 17: Design Specification for Review Classification

Tier 1: Data Persistence Tier
The Data Persistence Tier comprises of the data collection. The data on which this research is based on is collected from Kaggle[4]. This data is then fetched to a Jupyter Notebook for further detailed analysis.

Tier 2: Business Logic Tier
In the Business Logic Tier, the collected data is cleaned by handling the missing values followed by the exploratory data analysis. The customers are then segmented into 7 segments. Subsequently, the new time-based and distance-based features were engineered in order to enhance the performance of the machine learning model. The data is then oversampled and a stratified train-test split was performed on the data to feed the training data to a machine learning model. Machine Learning models such as Random Forest classifier, LGBM classifier, and AdaBoost classifier are used for the binary classification of review score. The implemented models are then assessed using the performance metrics such as accuracy and F1-score to find the best performing model.

5 Implementation

This section concentrates upon the machine learning models implemented to categorise the reviews as positive or negative. According to the interpretation of this study, a rating of 3 or higher is considered as ”positive”, whereas a rating of 3 or lower is considered as ”negative”, indicating that the quality of the product or service did not meet the receiver’s expectations. A binary classification is performed to classify the reviews using Random

[4] `https://www.kaggle.com/olistbr/brazilian-ecommerce`

Forest Classifier, Light Gradient Boosting Model (LGBM), and AdaBoost model. Out of these three models, the model with the best performance is selected.

5.1 Handling Class Imbalance

As observed in the Exploratory Data Analysis, the data is highly imbalanced (79% posiive and 21% negative). A model trained on an imbalanced dataset is likely to differ in the capability of predicting specific class. Oversampling the data is the most common practice for handling the imbalanced data. For this purpose, Synthetic Minority Oversampling Technique (SMOTE) and Random Sampling is used. The distribution of features when randomly oversampled was similar to the original distribution of features as observed in the original dataset. However, with SMOTE oversampling, the distribution is slightly deviated as compared to that of the original data. Therefore, randomly oversampled data is used for training the models.

The randomly oversampled data is used for splitting it into the training(80%) and testing(20%) datasets. The stratified splitting of the data is performed to ensure that the train and test sets contain about the same proportion of samples from each target class as the whole set.

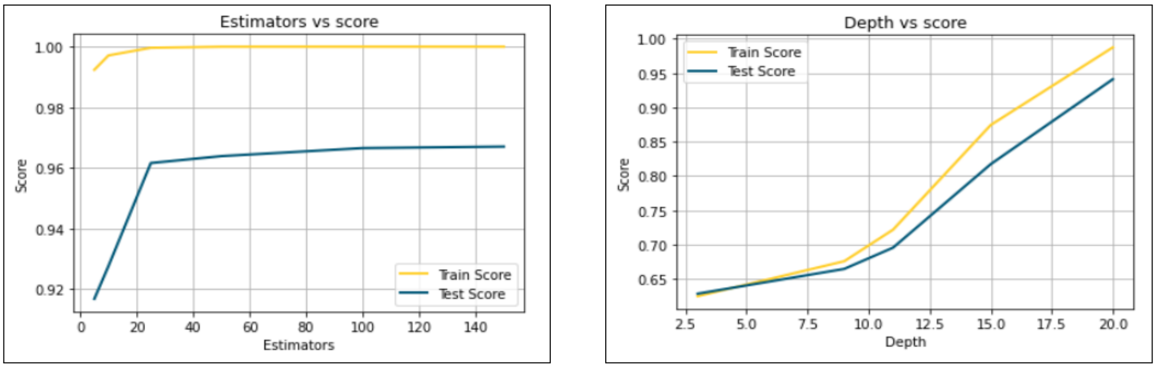
5.2 Data Preparation

The significant features that were engineered using the existing variables are used in model training. However, the attributes on top of which these features were engineered are discarded from the dataset along with other insignificant attributes. Also, the categorical variables such as "payment type", "order status", "product category name", and "RFM Level" are transformed in the numeric values to train the model using LabelEncoder.

5.3 Machine Learning Models

5.3.1 Random Forest

The Random Forest classifier using 10-fold randomized search cross-validation and 5 iterations is implemented on the randomly oversampled data to train the model. The RandomizedSearchCV is used to achieve the best hyperparameters in order to enhance the performance of the model. The basic hyperparameters of a Random Forest classifier are n_estimators, max_depth, min_samples_split, min_samples_leaf, and criterion. A list of potential values for each of the hyperparameters is parsed through a baseline Random Forest classifier with multiple iterations.



(a) Estimators Vs Score in RF Model

(b) Depth Vs Score in RF Model

Figure 18: Estimators and Depth in Random Forest Model

The Figure 18 shows the graphical representation of n_estimators and max_depth values parsed to achieve the best results. It was found that the best performance of the Random Forest classification model is achieved for 100 estimators and the default value of max_depth i.e., "None". Each iteration takes into account the values within the specified list of hyperparameters and returns the mean of F1 Macro score of train and test. RandomizedSearchCV is an approach to choose a subset from a combination of hyperparameter values and run against a number of cross validation and iterations.[5]. This research used a RandomizedSearchCV for 10 fold cross validation and 5 iterations. The best parameters obtained are 100 estimators, max_depth = None, min_sample_split = 4, min_samples_leaf = 3 and "gini" criterion.

5.3.2 Light Gradient Boosting Model (LightGBM)

Based on the literature survey carried out, the LGBM is known for its effectiveness and faster execution. The basic hyperparameters for LightGBM n_estimators, subsample, max_depth, learning_rate, colsample_bytree. The list of hyperparameters is parsed using RandomizedSearchCV for obtain the best parameters for the model. The best parameters obtained for LightGBM are subsample = 0.3, n_estimators = 1500, max_depth = 10, learning_rate = 0.15 and colsample_bytree = 0.3. The model is then trained using these parameters to achieve best results.

5.3.3 AdaBoost Classifier

An AdaBoost classifier is a meta-estimator that starts by training a classifier on the original dataset and then fits further replicas of the classifier on the same dataset with the weights of erroneously classified instances changed such that future classifiers concentrate more on difficult situations.[6] The best hyperparameters obtained for this model are n_estimators = 250, learning_rate = 0.1 and algorithm = 'SAMME.R'.

[5] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html`

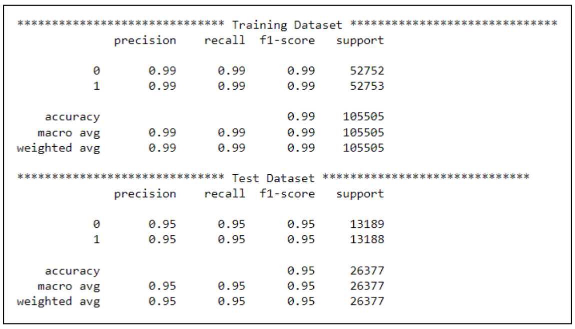
[6] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html`

6 Evaluation

This section of the research comprises of the evaluation of machine learning models implemented to identify the best suited model. The performance metrics used for evaluating the models are accuracy and F1-score based on the confusion matrix and classification report generated. All the machine learning models were utilized the same training and testing data.

6.1 Random Forest

The Random Forest model used for the binary classification of reviews was trained on the best parameters obtained as stated in the above section.



```
****************************** Training Dataset ******************************
              precision    recall  f1-score   support

           0       0.99      0.99      0.99     52752
           1       0.99      0.99      0.99     52753

    accuracy                           0.99    105505
   macro avg       0.99      0.99      0.99    105505
weighted avg       0.99      0.99      0.99    105505

****************************** Test Dataset ******************************
              precision    recall  f1-score   support

           0       0.95      0.95      0.95     13189
           1       0.95      0.95      0.95     13188

    accuracy                           0.95     26377
   macro avg       0.95      0.95      0.95     26377
weighted avg       0.95      0.95      0.95     26377
```

Figure 19: Classification Report Random Forest

The classification report in the Figure 19 shows that the overall accuracy of the test data using Random Forest model is 95% and the F1-score for positive and negative review class is 0.95.

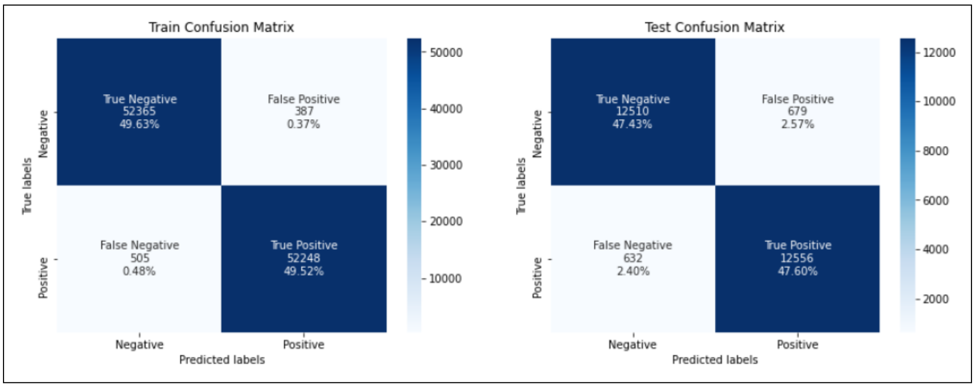


Figure 20: Confusion Matrix Random Forest

The confusion matrix in Figure 20 depicts true labels against the predicted labels. The random forest classifier was correctly able to predict 47.43% as negative reviews(TN) and 47.6% as positive reviews(TP). Performance of random forest model was best of all the models.

6.2 LGBM

The overall test accuracy for the LGBM model is 90% and the F1-score metric for both the positive and negative classes is 0.90.

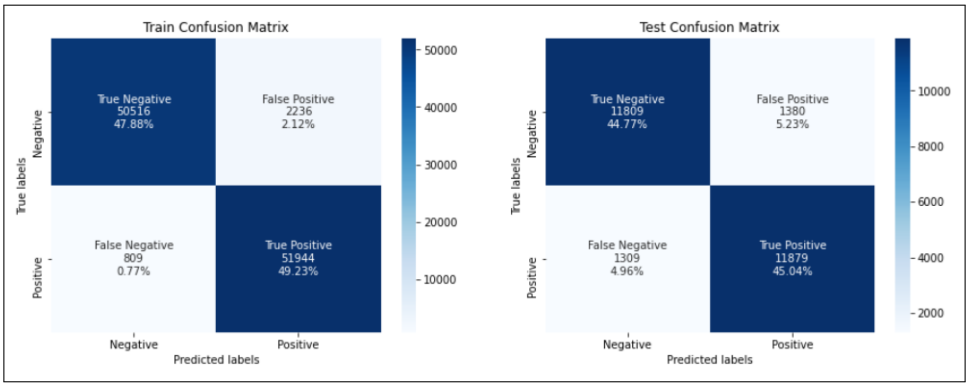


Figure 21: Confusion Matrix LGBM

The confusion matrix in Figure 21 depicts true labels against the predicted labels. The LGBM classifier was correctly able to predict 44.77% as negative reviews(TN) and 45.04% as positive reviews(TP).

6.3 AdaBoost Classifier

The overall test accuracy for the AdaBoost classification model is 66% and the F1-score metric for negative review class is 0.58 and positive class is 0.72.

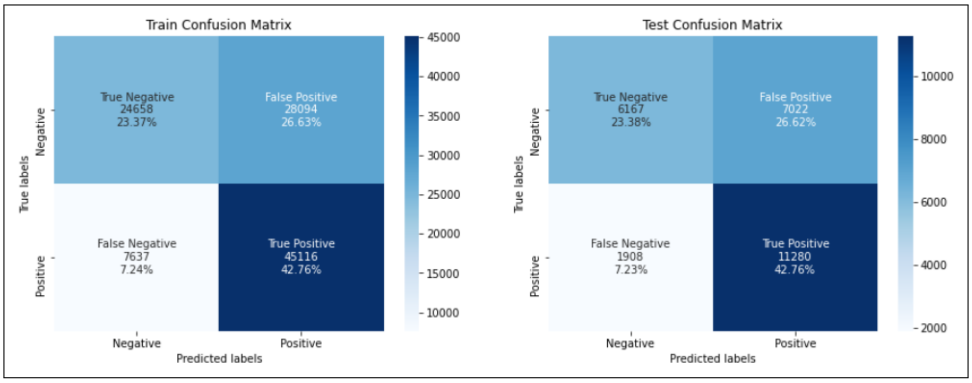


Figure 22: Confusion Matrix AdaBoost Classifier

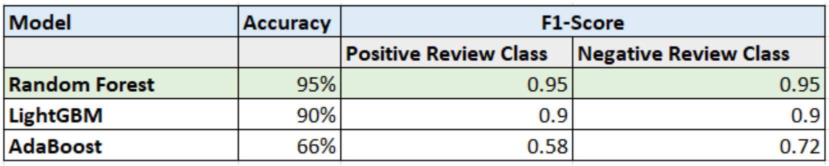
The confusion matrix in Figure 22 depicts true labels against the predicted labels. The AdaBoost classifier was correctly able to predict 23.38% as negative reviews(TN) and 42.76% as positive reviews(TP).

6.4 Summary

Sections 5 and 6 provide comprehensive solutions to the research question raised in Section 1. The models performed well, since numerous studies have shown that these models function well for classification issues. In this research, the Random Forest model performed well and outperformed all other models. As a result, Random Forest model is chosen for the classification problem.

From the top performing product categories, ”la_cuisine” is identified as the qualitative product category as 100% of the reviews submitted for this particular product category were positive. The best performing model trained using Random Forest classifier is used to test the data for this product category to verify if the model is able to categorise all the reviews as positive. It was observed that this model was able to classify all the reviews as positive with 100% accuracy.

The table below depicts the comparison between performance of the implemented machine learning models. It was found that that the Random Forest classifier outperforms LGBM and AdaBoost classifier in terms of accuracy and F1-score.



Model	Accuracy	F1-Score	
		Positive Review Class	Negative Review Class
Random Forest	95%	0.95	0.95
LightGBM	90%	0.9	0.9
AdaBoost	66%	0.58	0.72

Comparison between the Machine Learning Models

7 Conclusion and Future Work

The research presented the detailed analysis of data of an e-commerce marketplace integrator, Olist, situated in Brazil. In particular, the approach for segmenting the customers into 7 different segments based on their purchasing patterns (recency, frequency, and monetary value) and a binary classification approach to classify the review score as positive and negative is implemented using machine learning models. In order to perform the binary classification, review score greater than 3 were classified as positive reviews and ratings less than or equal to 3 are classified as negative reviews. In customer segmentation, it was found that most of the customers belong to the ”Potential” segment. In order to enhance the performance of the machine learning model, new time-based and distance-based features were created and analysed to check if the newly created attributes are correlated with the target variable. In order to handle the class imbalance, the data is randomly oversampled and is divided into the stratified train and test data. The training data was then fed to the Random Forest, LightGBM and AdaBoost classifiers for training. These models were assessed on the evaluation metrics such as accuracy and F1-score. The Random Forest model achieved the best results out of the three machine learning models implemented with 95% overall accuracy and the F1-score of 0.95 for both the positive and negative classes. This random forest model was also used for testing the product category ”la_cuisine” with 100% positive reviews to verify the performance of the model. The model was correctly able to classify all the reviews of this product category as positive.

As a part of future work, the number of features used for training the model could be reduced. Also, the deep learning techniques could be implemented for classification of the reviews. Furthermore, the usage of a personalized Recommendation Systems with either user-based or item-based collaborative filtering is a viable option.

Acknowledgement

I would like to express my sincere gratitude to Prof. Christian Horn for his unwavering support and guidance throughout the research, and thoroughly answering my questions and concerns. His assistance and counsel empowered me to tackle the concerns highlighted by my research.

References

Aravindan, S. and Ekbal, A. (2014). Feature extraction and opinion mining in online product reviews, *2014 International Conference on Information Technology*, pp. 94–99.

Chang, E.-C., Huang, S.-C. and Wu, H.-H. (2010). Using k-means method and spectral clustering technique in an outfitter's value analysis, *Quality Quantity* **44**(4): 807–815.

Chiliya, N., Herbst, G. A., Roberts-Lombard, M. and Bag, P. (2009). The impact of marketing strategies on profitability of small grocery shops in south african townships.

Dietterich, T. G. (2004). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, *Machine Learning* **40**: 139–157.

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

Hung, C. and Tsai, C.-F. (2008). Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand, *Expert Systems with Applications* **34**: 780–787.

Kaymak, U. (2001). Fuzzy target selection using rfm variables, *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*, Vol. 2, pp. 1038–1043 vol.2.

Kirasich, K. (2018). Random forest vs logistic regression: Binary classification for heterogeneous datasets, **1**(3): 25.

Kotler, P. and Keller, K. (2006). Marketing management, *Upper Saddle River, New Yersey* .

Leveraging non-respondent data in customer satisfaction modeling (2021). *Journal of Business Research* **135**: 112–126.

Lumsden, S.-A., Beldona, S. and Morrison, A. M. (2008). Customer value in an all-inclusive travel vacation club: An application of the rfm framework, *Journal of Hospitality Marketing Management* **16**: 270–285.

Machado, M. R., Karray, S. and de Sousa, I. T. (2019). Lightgbm: an effective decision tree gradient boosting method to predict customer loyalty in the finance industry, *2019 14th International Conference on Computer Science Education (ICCSE)*, pp. 1111–1116.

Marcus, C. (1998). A practical yet meaningful approach to customer segmentation, *Journal of Consumer Marketing* **15**: 494–504.

Maryani, I., Riana, D., Astuti, R. D., Ishaq, A., Sutrisno and Pratama, E. A. (2018). Customer segmentation based on rfm model and clustering techniques with k-means algorithm, *2018 Third International Conference on Informatics and Computing (ICIC)*, pp. 1–6.

Mccarty, J. and Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of rfm, chaid, and logistic regression, *Journal of Business Research* **60**: 656–662.

Miglautsch, J. (2000). Thoughts on rfm scoring, *Journal of Database Marketing & Customer Strategy Management* **8**: 67–72.

Monett, D. and Stolte, H. (2016). Predicting star ratings based on annotated reviews of mobile apps, *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 421–428.

Reddy, C. S. C., Kumar, K. U., Keshav, J. D., Prasad, B. R. and Agarwal, S. (2017). Prediction of star ratings from online reviews, *TENCON 2017 - 2017 IEEE Region 10 Conference*, pp. 1857–1861.

Sánchez-Franco, M. J., Navarro-García, A. and Rondán-Cataluña, F. J. (2019). A naive bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services, *Journal of Business Research* **101**: 499–506.

Tama, B. A. (2010). Penetapan strategi penjualan menggunakan association rules dalam konteks crm.

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

Wang, C.-H. (2010). Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques, *Expert Systems with Applications* **37**(12): 8395–8400.

Wardhani, N. W. S., Rochayani, M. Y., Iriany, A., Sulistyono, A. D. and Lestantyo, P. (2019). Cross-validation metrics for evaluating classification performance on imbalanced data, *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pp. 14–18.

Wei, J.-T., Lin, S.-Y. and Wu, H.-H. (2010). A review of the application of rfm model, *African Journal of Business Management December Special Review* **4**: 4199–4206.

Wu, H.-H. and Pan, W.-R. (2009). An integrated approach of kano model and anova technique in market segmentation — a case of a coach company, *Journal of Statistics and Management Systems* **12**: 679 – 691.

Yeh, I.-C., Yang, K.-J. and Ting, T.-M. (2009). Knowledge discovery on rfm model using bernoulli sequence, *Expert Syst. Appl.* **36**: 5866–5871.

Yu, M., Xue, M. and Ouyang, W. (n.d.). Restaurants review star prediction for yelp dataset, p. 7.

Zeinalizadeh, N., Shojaie, A. A. and Shariatmadari, M. (2015). Modeling and analysis of bank customer satisfaction using neural networks approach, **33**(6): 717–732.